

The Heliophysics Data and Model Consortium: Enabling Scientific Discovery with the NASA Heliophysics System Observatory

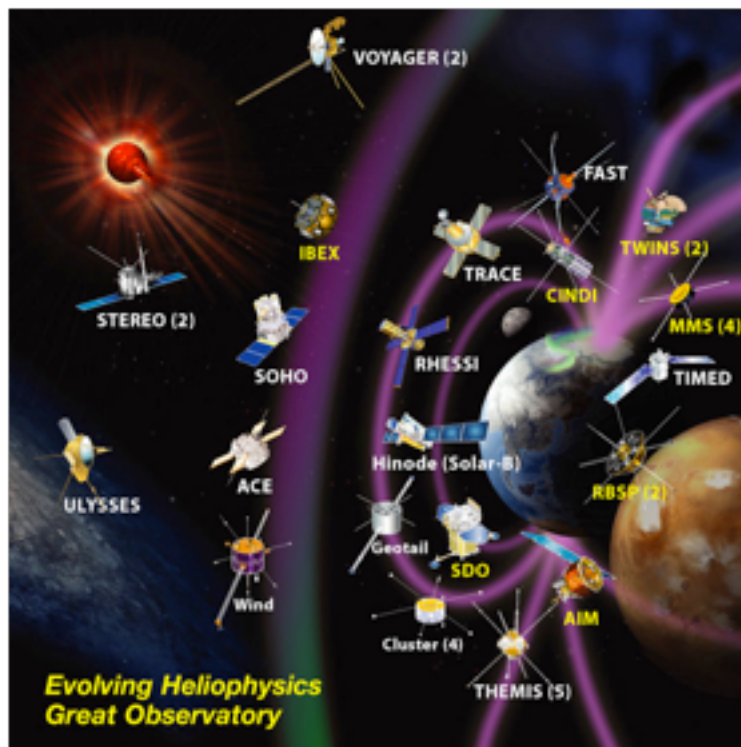
*Prepared for the NASA Heliophysics Data and Modeling Center
Senior Review, June-July 2009*

A Proposal for FY10-FY14 Activities

D. Aaron Roberts (NASA GSFC), Project Scientist
and the HDMC Team

Submitted 15 June 2009

E Pluribus Unum



The Heliophysics Data and Model Consortium:
Enabling Scientific Discovery with the NASA Heliophysics System Observatory

TABLE OF CONTENTS

Introduction and Overview	1
<i>Mission, Requirements, and Expected Accomplishments</i>	<i>1</i>
<i>Plan of the Proposal</i>	<i>5</i>
A Brief History	6
Science Goals of the HDMC	8
<i>Use Cases</i>	<i>9</i>
<i>Interdisciplinary Reality: Real overlaps and real problems</i>	<i>12</i>
Current Status	12
<i>Inventory.....</i>	<i>12</i>
<i>Virtual Observatories.....</i>	<i>13</i>
<i>Resident Archives.....</i>	<i>14</i>
<i>Data Upgrades.....</i>	<i>14</i>
<i>Value-Added Services.....</i>	<i>15</i>
Proposed Work.....	15
(1) <i>Inventory and IDs</i>	<i>15</i>
(2) <i>Registry Services and Access</i>	<i>16</i>
(3) <i>Finding and Accessing Resources</i>	<i>17</i>
(4) <i>Browsing and Visualization</i>	<i>17</i>
(5) <i>Solving the “Formats Problem” and “Dataset Runs-on-Request”</i>	<i>18</i>
(6) <i>Tool and Service Development</i>	<i>19</i>
(7) <i>SPASE Descriptions.....</i>	<i>19</i>
(8) <i>Data Quantity and Quality</i>	<i>21</i>
<i>Working Groups</i>	<i>22</i>
Roles of Specific HDMC Components: Individual VxO status and plans	22
<i>The Original: VSO.....</i>	<i>22</i>
<i>Next set: VHO, VMO (U, G), ViRBO, VITMO</i>	<i>23</i>
<i>Newer: VEPO</i>	<i>25</i>
<i>Newest: VWO, VMR</i>	<i>25</i>
<i>Soon: SuperMAG.....</i>	<i>26</i>
<i>Related: VSTO, GAIA, CSSDC, HELIO</i>	<i>26</i>
Outreach, Evaluation, and Feedback	27
Plan of Work and Milestones	27
<i>Management Plan.....</i>	<i>27</i>
<i>Milestones.....</i>	<i>28</i>
Budget Details and Justification	29
<i>In-Guide.....</i>	<i>29</i>
<i>Optimal</i>	<i>30</i>
References (URLs and Acronyms)	31
<i>Other Acronyms</i>	<i>31</i>

Introduction and Overview

The goal of the Heliophysics Data Environment, defined as the collective set of data from the “Heliophysics System Observatory” fleet and related resources, is to enable science discovery. The Data Environment provides essential infrastructure supporting NASA’s Vision to understand “the Sun, the Heliosphere, and planetary environments as elements of a single interconnected system, one that contains dynamic space weather and evolves in response to solar, planetary, and interstellar conditions” [NASA *Sun-Solar System Connection Science and Technology Roadmap 2005-2035*]. *We have a rich set of data, but it is in a wide variety of places and formats, and available through a varied collection of interfaces. This proposal offers a plan to provide the right lightweight infrastructure to support the uniform discovery, access, and use of a comprehensive set of space and solar physics data resources.* We want to make obtaining a useable form of the necessary data the easiest part of any Heliophysics science exploration task. New science is enabled when much more time can be devoted to analysis than to data discovery and preparation. Using new datasets should be a matter of browsing and understanding their content; too often we never get to that point because the data are difficult to read or lack basic documentation. We propose to make such difficulties a thing of the past for Heliophysics.

Our key concept is that of integration: bringing together diverse things and making them useful with each other. The goal of integration will be enabled by ongoing support for the upgrading, serving, and preserving of data products; the provision of browser and computer interfaces for "one-stop shopping" for data based on uniform terminology; the completion and maintenance of a comprehensive product inventory; and the development of basic services that allow the user to browse and use any datasets. Integration includes the idea of standardization through the use of uniform descriptions for spacecraft, observatories, instruments and data. It includes the idea of advanced searches that locate and gather data (or pointers to data) from many distributed providers, and it means simplifying or enhancing interfaces to many providers and making their data searchable with one interface. Integration includes providing tools that will allow the use of diversely formatted data, making it easy to perform side-by-side comparisons of diverse, distributed data products. We want to make various seemingly simple tasks, such as comparing two or more related science measurements, actually simple to execute. Ultimately, integration means developing standard tools for mapping between data sets in different regions of the nonlinear dynamical Sun-Earth and Sun-Planetary environments that are the domain of Heliophysics. The resulting global views cannot help but trigger new types of questions and new ways of doing science enabled by the Heliophysics Data Environment.

The Heliophysics Data and Model Consortium (HDMC) was initiated on October 1, 2008 to formalize this integration as part of the overall Heliophysics infrastructure. While formally a NASA organization, with specific NASA-funded elements that are being reviewed here, this consortium is international since the data needed for space physics research are distributed around the world. The HDMC was built on a number of years of grass-roots efforts and, as discussed in the brief historical sketch below, it was born of the need to provide coherence, effective coordination, and a light management framework for a number of new elements of the Heliophysics Data Environment.

Mission, Requirements, and Expected Accomplishments

The Heliophysics Science Data Management Policy (see <http://hpde.gsfc.nasa.gov>) defines the roles and responsibilities of the HDMC in Appendix G. The Data Policy is intended to foster the greatest possible scientific use of Heliophysics data resources based on the principles of providing open, scientifically useful data, and involving scientists in all aspects of the data lifecycle. Specifically, the Policy states that:

The Mission of the HDMC is to facilitate Heliophysics research, both local and global, by providing open, easy, uniform, scientifically meaningful access to a comprehensive set of relevant resources (data, models, tools, and documentation) as quickly as possible from the time each is created, and for as long as each resource is deemed useful by the Heliophysics science community.

An analogy may make our aims clearer: We plan to produce a “library for Heliophysics data” in which all books (data) are stored, known, categorized, and searchable in an online catalogue. It would be better than a standard library, however, because all books (data) could be checked out as many times as needed, in the desired language, and with ways to determine which parts of the books are relevant through summaries (browse plots and images), specialized catalogues, and targeted searches of condensed versions (modest resolution uniform datasets).

To carry out this mission, the HDMC is required in the Data Policy, Appendix G, to accomplish the following objectives (which we take to be our “Level 1 Requirements”):

1. Define, implement, and maintain a data environment that enables access to a comprehensive set of distributed heliophysics resources using uniform interfaces and standards by:
 - (a) creating and maintaining an inventory with a basic registry and easy access to resources;
 - (b) developing and maintaining discipline specific search and access tools;
 - (c) developing and maintaining interoperability standards; and
 - (d) monitoring the continued utility of a core set of formats (HDF, CDF, FITS, ASCII).
2. Manage specific post-mission datasets by
 - (a) maintaining approved Resident Archives for preserving and serving post-mission data, and
 - (b) upgrading legacy datasets for accuracy, completeness, easy access, and utility.

Given the requirements for the HDMC, we propose to:

- (1) Complete a comprehensive, accessible **Inventory** of Heliophysics data and related resources, with links to the data through a single portal that also links to more capable portals;
- (2) Provide **discipline specific portals** to Heliophysics resources (“VxOs”) that add value by providing easy-to-use interfaces and search tools based on events, positions, and other criteria;
- (3) Provide basic, generic tools that allow the **reading and display of data in any format**, thus allowing browsing of all resources and a uniform output of the data;
- (4) Develop and implement a “**Dataset Runs on Request**” service that will allow a user to request both small and large runs to produce subsetting, merged, interpolated, and/or averaged data files in a desired format and granularity;
- (5) Continue to provide **opportunities for the development of services** linked to Archives and Virtual Observatories.
- (6) Complete and maintain **uniform** (“SPASE” plus linked documentation) **descriptions of the inventoried resources** that are adequate to find, access, and use the resources for research;
- (7) Develop and implement a **uniform language** that makes it easy to implement machine-to-machine requests for metadata and data (a “SPASE Query Language”), using a variety of applications as the interface;
- (8) Maintain the serving of mission data after missions end (via **Resident Archives**), and assure that useful legacy data are transitioned to Final Archives; and
- (9) Continue to **upgrade legacy data** to make it easily used and served.

To implement these tasks, the HDMC proposes to continue on the path that has been initiated by specifically funded NASA Research Announcement (NRA) selections (to continue as HDMC-funded) and the formation of the Data Policy. Table 1 gives an overview of how this will be done, and Fig. 1 gives a schematic view of the overall system architecture. The HDMC will consist of Resident Archives; data recovery and upgrade projects; discipline specific Virtual Observatories (VxOs, where “x” stands for, e.g., “Solar” or “Magnetosphere”) and associated value added services; and the SPASE consortium that is responsible for the HP data model. These will be described further below. Decisions on the direction of the project are to be made by an Implementation Working Group, led by the HDMC Project Scientist, that will consist of representatives of the constituent HDMC groups (VxOs, SPASE, etc.)

and the NASA Data and Computing Centers. A major focus of this proposal is a plan to take existing VxOs to the next level in terms of completeness of data holdings, sophistication of search capabilities, and usefulness in data exploration and analysis.

The constituents of the HDMC were (and would continue to be) formed as the result of successful peer-reviewed proposals, with the subsequent evolution guided by community input from the Heliophysics Data and Computing Working group as well as more directly through the outreach efforts of each HDMC subgroup and through periodic formal Senior Reviews. The HDMC proposes to continue to work with national and international partners and EPO groups to maximize impact and utility. It will report to the NASA Headquarters Heliophysics Data Environment Program Manager. There would continue to be a NASA NRA for Data Upgrades (linked to the VxOs and RAs or Data Centers), initiating RAs, and initiating services (linked to VxOs and/or archives).

In addition to providing scientific leadership, the HDMC Project Scientist is the liaison for the HDMC community to HQ, and maintains an HPDE Web site with overviews of the Data Environment, current events of interest, and links to the HDMC components to keep the community informed of progress on HPDE issues.

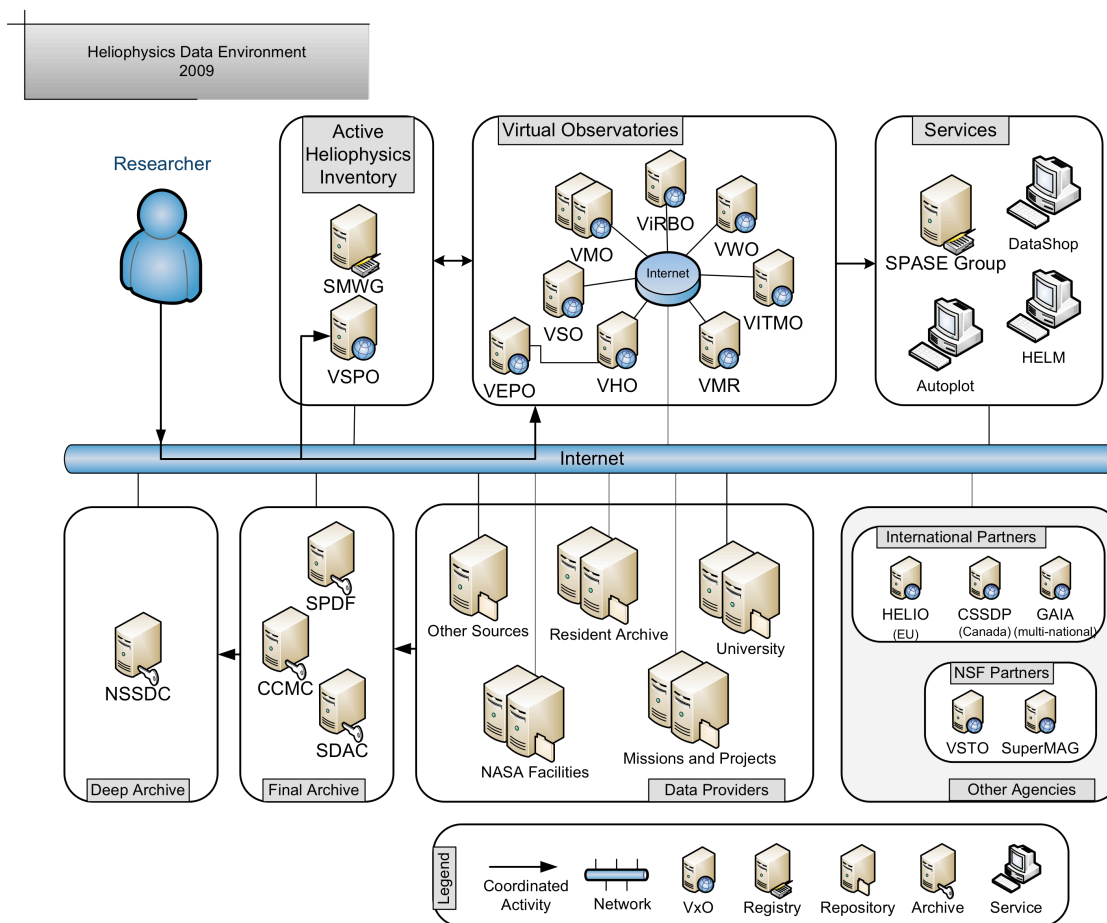


Fig. 1. The main components of the Heliophysics Data Environment. **The HDMC is directly responsible for the VxOs, the related Services, Resident Archives, and (with SPDF/VSPO) the Inventory.** Users access archives directly or through integrated portals (Virtual Observatories). The “Virtual Space Physics Observatory” (VSPO) provides routes to the full range of data and services, often by using VxO services. The VxOs provide access to services such as browsing of a wide range of resources and using event and feature lists to find relevant data. Final Archives are data providers, and the CCMC provides modeling resources as well. The NSSDC primarily serves as a backup archive for archives. International and Interagency partners provide both alternate integrated views of data and access to a wider range of data and services

Relationship of the HDMC to NASA Heliophysics Data Centers

The "Data Centers" in this review are the Space Physics Data Facility (SPDF), the Solar Data Analysis Center (SDAC), HDMC, and, in a somewhat different capacity, the National Space Science Data Center (NSSDC, which covers more than Heliophysics). All four are independent entities with separate budgets and proposals, but they form a coordinated whole along with the Community Coordinated Modeling Center, which is not being reviewed at this time. This section discusses the functions of the non-HDMC components and the relationships between all the components, as defined in the Data Policy. *The set of functions these groups cover, along with Mission data systems, constitute a complete data environment, providing for the tracking of the status and whereabouts of Heliophysics data and related resources; the means to find, access, and use data and models; the levels of archives needed to assure continuity of data availability; routes to recovering important legacy resources; and the means to keep data safe throughout its lifetime (see Fig. 2).*

Requirement (Data Policy Appendix G)	Specific tasks (keyed to above list)	Group(s) involved	Comments
Create and maintain a comprehensive inventory	Assure that all available Heliophysics data products are registered in a uniform way and available through one portal; Task (1).	SPDF via VSPO; VxOs, Data Centers, Missions	SPDF will maintain the "active inventory" but the content will be developed by many groups, primarily VxOs and Data Centers.
Develop and maintain discipline specific access tools	Provide discipline-specific data portals that allow users to browse, search for, and access data to efficiently accomplish research goals; Tasks (2–5).	Primarily VxOs working with Missions and Data Centers; also Value added service groups.	VxOs maintained as HDMC infrastructure; they develop core services (browse plots, generic data processing) with other services developed through NRA proposals.
Develop and maintain interoperability standards	Maintain and develop community-wide standards for description of products and the interaction of services; Tasks (6, 7).	SPASE group; VxOs; interagency and international groups	Data Model maintained by SPASE; SPASE-QL developed by HDMC; interagency and international cooperation desirable.
Preserve and serve post-mission data	Assure that data transition to post-mission and post-RA phases with utility maintained; Task (8).	Resident Archives; HDMC and Final Archive oversight and collaboration	RAs formed by proposal to NRA; maintained as HDMC infrastructure, continued utility determined in Senior Reviews.
Upgrade legacy data	Continually bring useful legacy data into more useful states (better quality, format, access); Task (9).	Data Upgrade proposal groups; VxO and Data Center assistance	Mostly achieved through NRA grants, but small tasks done by VxOs and Data Centers.

Table 1: Flow-down chart from requirements to tasks and the groups involved. Monitoring the health of essential formats is part of maintaining interoperability standards. There has been and will continue to be an entry in the NASA ROSES call for proposals that targets the NRA work in this table. The funding for the NRA is part of the HDMC budget.

SPDF is the primary Heliophysics Final Archive for Space Physics data, which is served through CDAWeb and other Web portals and ftp sites. Through specific agreements, it also serves data for a number of active missions, which makes final archiving easy. SPDF develops and maintains a variety of essential services, notably SSCWeb (spacecraft orbits and related services), and OMNI (an extensive database of 1 AU solar-wind and Sun/Earth activity index data with associated services). SPDF developed and actively supports the CDF format that is in very common

use by missions, VxOs, and the CCMC. Most of the SPDF capabilities and data are available to VxOs through Web Services. The Virtual Space Physics Observatory (VSPO), initially funded through various NRA grants and which is becoming the “face of the Heliophysics Inventory,” became part of the SPDF in the last Senior Review. SPDF has the computer expertise to maintain and upgrade VSPO, and the SPDF has a central role as a Final Archive that makes it appropriate as the keeper and maintainer of a general Heliophysics Inventory for all resources, not just those served by SPDF.

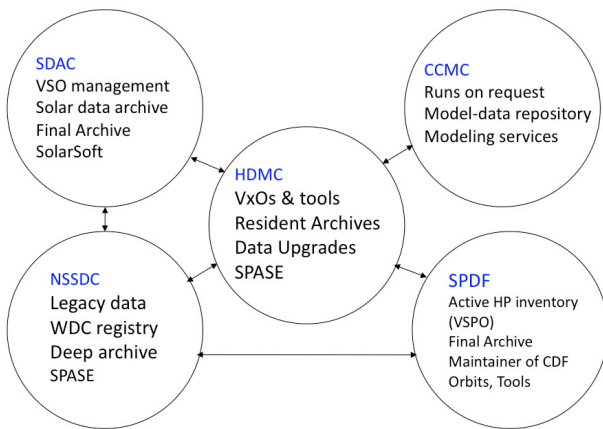


Fig. 2: Heliophysics Data and Computing Centers that together with Mission data systems form a complete Data Environment.

SDAC plays a role similar to SPDF but for the solar component of Heliophysics data. It works with missions to serve large amounts of solar data (e.g., STEREO and SOHO), and will play an essential role in the serving of Solar Dynamic Observatory data. SDAC will serve as the Final Archive for Solar data, although in some cases it will be responsible for assuring the data are well archived rather than holding it physically (e.g., SDO’s large data volume may be better handled in other ways). SDAC is also responsible for such essential tools as the SolarSoft analysis package, and, partly through HDMC projects, this and other analysis packages are becoming more widely useful in the VxO context. The Virtual Solar Observatory (VSO) is part of SDAC, and although at one point it was thought it should become part of the general group of VxOs, the previous Senior Review of Data Centers endorsed the idea that SDAC should

keep VSO. We decided in this round that this was working well, so we should not change it. VSO will collaborate as a VxO with the others, producing, for example, SPASE descriptions of data and linking to the others. (VSPO has such links already, along with basic SPASE descriptions of solar products.) The SPDF and SDAC, as Heliophysics Archives, will work closely with the HDMC to assure a seamless Data Environment, with each group performing complementary functions.

The NSSDC formerly included the SPDF as a component, but NASA decided to separate active data serving functions (SPDF) and the long-term safekeeping functions (still called NSSDC). Thus, NSSDC now primarily serves as an “Archive for Archives” that provides backup and media-handling/refresh capabilities that the active archives do not have the resources to maintain. The NSSDC serves this function not only for Final Archives, but also for active missions such as RHESSI and others, as requested by the missions. The NSSDC is the historical archive of early space science data. As part of its curation function, to migrate data from legacy media, NSSDC has undertaken tasks to prioritize and then modernize both its analog and digital heliophysics holdings. The NSSDC, an originator of the SPASE idea, provides funding for core SPASE services. The primary SPASE funding supports a coordinating PI and a principal maintainer of the documents, schema, and tools, both as much-less-than-half-time positions. Also, NSSDC has a role as a World Data Center, and thus provides information on missions and instruments that is harvested by the VxOs. The NSSDC supports other NASA Science Mission Directorates, but this role is not relevant here.

The Community Coordinated Modeling Center, while not reviewed at this time, serves the Heliophysics community by providing modeling services and model data, and the HDMC is directly linked to the CCMC through one of the VxOs (the Virtual Model Repository). The incorporation of model output and services as a natural part of a researcher’s workflow is a long-term goal that is a natural part of the HDMC. The HDMC works closely with the CCMC, exploiting tools such as Kameleon that allow uniform model descriptions and storage and efficient data access.

Plan of the Proposal

Many of the tasks we will perform should seem obvious as requirements on the Heliophysics data system. As an explanation for why the HDMC is a new initiative, the next section will discuss the history of the Heliophysics Data Environment, indicating how we arrived at a situation where the components we now propose did not exist. The

short answer is that such plans have been made, but, for various reasons, did not fully come to fruition, although progress has been made. Following that, we will provide the science rationale for the project in more detail, giving specific examples. The element of the HDMC most urgently under review here is that of the discipline specific Virtual Observatories (VxOs, where “x” refers to a subfield such as “M” for Magnetosphere), because the initial grants for these projects will be running out within months. (Note that the Virtual Solar Observatory has already become an infrastructure project run through SDAC, and will primarily be part of that proposal.) Thus, these VxOs will be entering their “extended missions,” while the more recent VxOs and the Resident Archives are still in their “prime missions” under NRA funding, although they, too, will reach the end of their initial grants within the five-year purview of this proposal. Much of the rest of the proposal will focus on issues concerning the VxOs and related value-added services, the role and expected growth of Resident Archives, and the successes and expected continuing role of the Data Upgrade part of the project. The Heliophysics Data Inventory, while an HDMC project in terms of its scope, is being implemented by the SPDF since they have the required organization and expertise, so much (but not all) of the discussion of this aspect of the project will be in the SPDF proposal.

A Brief History

Almost thirty years ago, NASA began to come to grips with the evolution toward both larger and more comprehensive spacecraft-based datasets and (somewhat later) the increasing capability of the Internet to deliver these data to scientists. The initial set of principles was set out in the “CODMAC report” (see inset). The efforts divided fairly early on, with the Planetary and Astrophysics Data Systems each taking its own course, and the Space

The National Research Council's Committee on Data Management and Computation (CODMAC) list of principles for successful scientific data management: 1. Scientific Involvement in all aspects of space science missions. 2. Scientific Oversight of all scientific data-management activities. 3. Data Availability - Validated data should be made available to the scientific community in a timely manner. They should include appropriate ancillary data, and complete documentation. 4. Facilities - A proper balance between cost and scientific productivity should be maintained. 5. Software - Transportable, well-documented software should be available to process and analyze the data. 6. Scientific Data Storage - The data should be preserved in retrievable form. 7. Data System Funding - Adequate data funding should be made available at the outset of missions and protected from overruns. (National Academy Press, 1982)

Physics Data System (the term then included Solar Physics) working separately. Unlike in the other cases, the Space Physics cats were not so easily herded, and, due to various reasons including a community perception that money for a data system was less money for science, this effort did not give rise to a coherent approach. Nonetheless, the ideas put forward in those early efforts were nearly the same as our current thinking, and the HDMC may be seen as part of the natural continuation of those efforts. See <http://hpde.gsfc.nasa.gov> for

documents and further information.

Figure 3 shows (viewgraph!) slides from the first formal meeting of the SPDS Steering Committee in 1990. The key issues are front and center, and largely derive from CODMAC: Scientists want to do research with complete, well-documented sets of data that are easy to find, access, and use; reliably stored; and delivered in convenient formats. We think of the emphasis on multimission studies as relatively new, but this was the ISTP era, with a fleet of a half-dozen or so craft setting out to solve the problems of magnetospheric physics. Pairs of Pioneer, Helios, and Voyager spacecraft, combined with ISEE-3 and IMP-8 near the Earth, had already provided multipoint measurements of the heliosphere for over a decade, and similar examples can be found in other areas. As the SPDS effort faded, although with some effect on data systems, data from past missions tended to languish with no clear plan for their long-term use and preservation beyond “send something to NSSDC.” The ISTP efforts did institute a Coordinated Data Handling Facility that provided data to the teams and made considerable efforts to document events in the data streams. The ISTP-based CDAWeb at GSFC was making strides in providing easy access, but many of the steps, such as obtaining adequate, uniform data descriptions, were difficult. There was an increasing acceptance of the open data policy that NASA was insisting on: public money implies public data. Many missions and more integrated Internet-based data systems, such as that at UCLA, were providing increasingly able access to data.

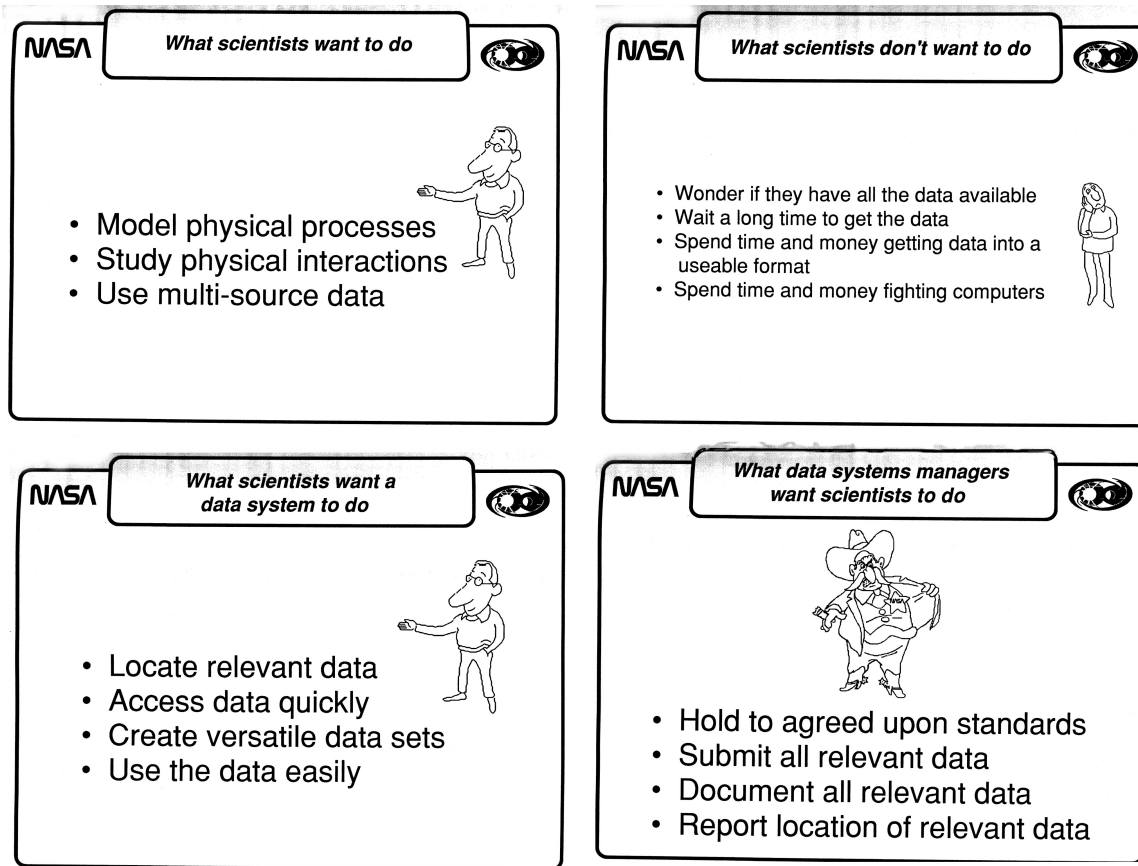


Fig. 3: Slides from the first presentation to the first Space Physics Data System Steering Committee meeting in October, 1990. The goals of the HDMC reflect the decades-long constancy of the needs of scientists and the data systems they will use.

The initial plan for the SPDS was to use the NASA Master Directory—an inventory of all data products—as a central means of identifying datasets and how to retrieve them; subsequently there would be a more automated but more expensive distributed system of data access to many data nodes. By 1998, the “River Bend Workshop” provided a detailed recommendation for a somewhat more centralized but basically similar approach. While some useful projects resulted from these efforts, they were never formalized. Meanwhile, Heliophysics increasingly enforced the open data policy (CODMAC Principle 3) through a data access component of the mission Senior Reviews. Each subsequent review asked for somewhat more, and produced more and better data from the missions. (See http://hpde.gsfc.nasa.gov/hpde_background.html for documents and more details.) In the same time frame, an eye-opening report came from the Living with a Star (LWS) Data Environment (when this was seen as a separate entity) that indicated a number of problems. A major one was that there was no way researchers in any subfield except perhaps solar physics could find out what data resources were available to them because there was no systematic tracking of such resources—there was clearly no longer a master directory equivalent.

As an important step toward a coherent Heliophysics data system, the MO&DA and Computation components of the Heliophysics Division at HQ established a Data and Computing Working Group (DCWG), comparable to the discipline-focused Heliophysics working groups, to provide community feedback. Starting in 2002, the DCWG heard presentations on Project Data Management Plans, Data Center operations, and on plans for “Virtual Observatories”—an idea that started in the astrophysics community and that was adopted by the Solar community in a Senior Review of Heliophysics Data Centers. The VOs represent the realization of the distributed data system envisioned in the early SPDS discussions, in which distributed resources are available simply and uniformly through single portals. The Living With a Star program was also making plans for a data system, and as it became clear that the LWS goals were Heliophysics-wide, these efforts were brought into the MO&DA line. A 2004 workshop further clarified the goals and structure of VOs, and became the basis for a call for proposals for “VxOs” consisting

of VOs for the “x” community (e.g. VMO for the Magnetosphere). Of crucial importance was the funding line that was established for VxOs, data restoration efforts, and other aspects of the data environment. Although NASA Heliophysics Data Center support declined somewhat, it stabilized and the total support for the data environment increased. The Space Physics Data Facility and the Solar Data Analysis Center continued to evolve as both active data repositories and parts of a distributed data environment.

To solidify the plan for the Data Environment, starting in 2006, NASA Heliophysics began developing a Science Data Management Policy based on the basic CODMAC ideas of open data and scientific involvement in all aspects of the data environment. Community input came from the DCWG and many other sources. This Policy, first released in July 2007, provides a blueprint for using existing data environment resources and new initiatives to realize the CODMAC and SPDS goals. It provides a comprehensive overview of the data environment. The first version of the document had some significant omissions, but the current version provides a more complete picture. One of the last elements added was the defining text for the HDMC as the coordinating umbrella primarily for the aspects of the data environment that are created by proposing to the NASA Research Announcement that includes VxOs, data restoration, and Resident Archives.

Thus we arrive at the present with a very different picture from what had existed until very recently. The community has become accustomed to the need for a consistent approach to the production and curation of data, and for easier means to perform research using many datasets. A Data Policy exists to provide coherence and coordination. Both new and existing missions have clearer guidance on how to make data more easily and openly available, and on what NASA expects for the near and long term. The HDMC goals outlined in the Data Policy and discussed above are still unrealized in a number of respects, but there are many successes to report, and the direction is now clear. An inventory of Heliophysics data and related resources has made considerable progress based on the Virtual Space Physics Observatory (to become the “Heliophysics Resource Gateway”), which started at the same time as VSO and has now been absorbed by the SPDF. This Inventory provides the realization of the simple initial data system envisioned by the SPDS in which a complete inventory provides access to all resources, although newer technology makes it much more capable than envisioned. Discipline specific Virtual Observatories have been started in all subfields, with the VSO leading the way to mature functionality; initial individual efforts are now becoming more coordinated. A language for describing Heliophysics resources (the “SPASE Data Model”) is now defined and being used after considerable grass-roots struggle. Many datasets that have been unavailable have been restored and others that were in danger have been preserved. The continued serving of post-mission data is assured through the funding of Resident Archives, and the Data Policy defines a clear path to preserving useful data products in Final Archives. This is the context, then, for this Senior Review, which provides an opportunity to set out specific goals and an implementation path to finally realize the goals implicit in the presentation in Fig. 3.

Science Goals of the HDMC

The HDMC is designed to facilitate Heliophysics science research, which is the ultimate justification of its existence. The provision of new and continuing datasets through Resident Archives and Data Upgrades are self-evident as being essential to Heliophysics endeavors, in that observational data form the core of any scientific conclusion. Each of these projects will be judged on its own merits as a contribution to the Data Environment through the NRA process. Resident Archives will be judged to be worth continuing through Senior Reviews such as this one, but in this case they are all only starting. However, the integrative function of the VxOs is new, and thus it is important to show how it can aid research. None of the VxOs except the VSO have been in existence long enough to have achieved completeness or to provide nearly the level of utility we expect to achieve in the next couple of years, so we do not yet have significant science literature citations to report, although we do have many presentations that address both software architecture and science uses of the various components of the HDMC (see <http://www.spase-group.org/biblio.jsp> and the references provided at many VxO sites). The examples below, mostly actual current use cases, illustrate the benefits. Note that the expectation of extended support for the HDMC will provide incentive to users to begin relying on the Virtual Observatory services more extensively, now that they know that these services will not go away after their initial NRA funding.

Based on our experience to date, we have a clearer answer to the basic question: What can virtual observatories and related services really do for us? In many discussions it is assumed that the goal of VxOs is the solving of global

problems involving many different resources, and this is indeed one part of the picture. However, probably more important is a new way of interacting with data that enables research of all sorts. Having multi-scale browse and numerical data easily found and displayed makes the data environment a powerful tool for investigation that relieves endless days of drudgery, thereby inviting investigations that would have been dropped as not practical. It becomes very easy to see the context of events, view overview plots to obtain rapid insights into the nature and applicability of particular datasets, go directly to the data for a particular instrument, find and download datasets from other spacecraft for an event or at higher resolution from a single instrument, obtain the data for all events of a particular kind in a specified time range, and myriad other tasks that are obviously useful but, until recently, not easily accomplished. Simple questions are now rapidly answered: Is this solar minimum atypical? What other spacecraft were in the magnetosphere when this substorm occurred? Are there magnetic field data at better than 10 second resolution in the heliosphere in 1978? Where can I get them? What are the persistent features in the X-ray corona during a given interval? When did a spacecraft measure mesospheric winds over my ground station? These and endless questions like them are now answered in seconds or minutes rather than days. Those who now use these new tools are beginning to see a fundamental change in their way of working. The change is similar to that brought about by the advent of online abstract search engines such as the NASA/ADS. There is no longer any reason not to know what the latest papers are on a given topic. The result of the HDMC project will be that this ease is repeated in the realm of data discovery, access, and use. The tools we are developing will not just be for those studying space weather, although they will benefit as well. Nearly everyone who uses Heliophysics data will find they are increasingly able to focus on the content and implications of data rather than where they are or how to change their format. What we are working on and proposing will bring on a quiet revolution not just in the ease of larger-scope projects, but also in the details of our daily "simple" interactions with data.

Use Cases

Characterizing solar wind discontinuities. A researcher made a claim that in the solar wind, nearly constant magnetic field intervals in which the components undergo a discontinuous change only occur close to the Sun (0.3 AU in Helios data). Using VSPO to find plots of outer heliospheric magnetic field data yielded Ulysses magnetic field browse plots, at the Ulysses site. Exploring for a few minutes there produced likely candidates with the desired characteristics. Returning to VSPO revealed 1-second resolution Ulysses magnetic field data at CDAWeb. A few more minutes work yielded plots of intervals that met the criteria on the magnitude and components of the field, thus providing counter-examples to the original statements. Total time for this new result: about an hour.

When the solar wind disappeared. On April 26-27 and May 10-12, 1999, unusually low solar wind densities produced a very weak Earth bow shock that moved out past 100 Re upstream of Earth and possibly all the way to L1. A number of space science articles were published in 2001 analyzing these unusual events that provided an excellent testing of the various magnetospheric models during extreme situations. At that time, it took over a week searching for other low-density solar wind events to establish the frequency of these unusual conditions. Today, with VHO, it took just 10 seconds of search time to find 100 days when Wind observed solar wind densities below $1/\text{cm}^3$ further than 50 Re upstream from Earth. Moreover, in each day, all such intervals to the nearest 10 minutes are individually called out.

Interplanetary shocks in the magnetosheath. Comparing observations of interplanetary shocks that were observed in both the undisturbed solar wind and in the Earth's magnetosheath allows the investigation of the nature of shock-shock (interplanetary shock – Earth's bow shock) interactions. Individual interplanetary shocks observed by Wind have been tabulated in a number of publically available catalogues (soon to be ingested into VHO). Specifying these shocks times as the search time window and specifying the region of interest as magnetosheath will in a few seconds provide all concurrent magnetosheath observations that are available. Using Autoplot, these magnetosheath intervals can be quickly reviewed looking for large magnetic field intensity increases. Thus in less than one day, the work of a month or so by more manual means can be reproduced.

Power spectral evolution in the solar wind. It has been found that the slope of the power spectrum of solar wind velocity at 1 AU is not the same as that of the magnetic field, thus questioning a basic assumption about solar wind turbulence. Since the spectrum is known to be evolving, the question arises what is happening farther out. A search in VSPO for plasma data in the outer heliosphere revealed that Voyager and Ulysses both had potentially useful data. Examining the product descriptions showed Voyager had higher resolution. The VSPO link to COHOWeb allowed a quick survey of hourly averages of Voyager data to find potential candidate regions with sufficient

coverage and properties appropriate for spectral analysis. Returning to VSPO revealed a source of the higher resolution plasma data at MIT; an ftp site provided easy access. It did take an hour or so to write an IDL routine to read the ASCII files, but then it was possible to perform the spectral analysis to determine that the slope of the velocity spectrum farther out in the heliosphere became the same as that for the magnetic field, thus changing the debate to being about why the magnetic and spectral evolution occur at different rates. This new result was found in the evening before the second day of a session at a conference. It is taking longer to write up the results than to do the research.

The temporal and spatial evolution of magnetospheric substorms. Slavin et al. [2002; JGR, DOI 10.1029/2000JA003501] examined simultaneous observations of earthward flow bursts and plasmoid ejection during magnetospheric substorms. In order to study the temporal and spatial evolution of the substorms, they searched for a radial alignment of spacecraft from geosynchronous orbit (substorm current wedge formation), to near tail (earthward flow bursts) and the deep tail (plasmoid ejection) during a period of time when observations of the upstream solar wind and auroral oval were available. They used GOES and Geotail as sources of measurement in the inner magnetosphere and near-tail, IMP 8 for the detection of plasmoid ejection in the deep tail, and Polar for auroral images. The IMP 8 data from June 1996 to October 1997 were searched for traveling compression region signatures and found 43. Next, they restricted the Geotail database to times when the spacecraft was located in the pre-midnight or midnight region of the near tail, i.e. $X > -15$ Re, and the Geotail-IMP8 separation in GSM Y was < 10 Re. These requirements were strictly satisfied for only two substorms that occurred early on 9 July 1997.

The above requirements were restated by using a VMO search interface for structured queries into the following set of conditions:

1. Time interval 1 June 1996 to 31 October 1997
2. IMP 8 magnetic field data located between $X_{\text{gsm}} < -8$ Re and $|Y_{\text{gsm}}| < 10$ Re
3. Geotail plasma and magnetic field data located between -15 Re $< X_{\text{gsm}} < -5$ Re and $|Y_{\text{gsm}}| < 10$ Re; and intervals with oscillating direction of the plasma flow V_x velocity component: $\min(V_x) < -200$ km/s and $\max(V_x) > 200$ km/s.
4. GOES 8/9 data at $X_{\text{gsm}} < 0$ Re
5. Solar wind and interplanetary magnetic field (IMF) data.

Note that the VMO query is not precisely the same as restrictions used by Slavin et al. The query has been adapted to suit the capabilities of the VMO. Also note that the search for the solar wind and IMF data was performed by the VHO based on the VMO-identified time intervals, thus demonstrating the inter-VxO communication and/or collaboration. The VMO/VHO identified 6 data files for download/review and specifically identified one of the events studied by Slavin et al. The second event, even though available in the identified files, was not marked by the VMO because the V_x oscillations were a bit less pronounced and would be easily detected if the condition number 3 above used a speed limit of 150 km/s, for example. However, the VMO search took only about 1-2 hours including both query preparation and processing while Slavin et al. Spent 10-100 hours identifying the events [Slavin, 2008 personal communication]. The VMO query could be quickly varied and rerun with many conditions, whereas the manual search would likely only be done once.

Acceleration of Energetic Particles in the Magnetosphere – A THEMIS Use Case. On March 1, 2008 the five THEMIS spacecraft were in conjunction in the near-Earth tail. Four of the spacecraft were aligned in the tail at $Y_{\text{GSM}} = 6$ RE and at X distances between -23 and -8 RE. THEMIS investigators noticed that the energetic particle fluxes (up to 500 keV) within the near-Earth plasma sheet ($x \sim -8$ to -10 RE) increased by orders of magnitude in about 2 minutes. The investigators want to understand the physics of this increase. Is it just boundary motion or does this represent an example of particle acceleration? If acceleration is responsible then what is the acceleration mechanism? What role does the ionosphere play during such events? On March 1 the THEMIS spacecraft provided excellent observations at a single local time. In addition to the THEMIS energetic particle data, the investigators need THEMIS plasma data and magnetic field data to address these questions. Although there are five THEMIS spacecraft, the magnetosphere is just too vast to be studied without using data from a large suite of spacecraft and observatories. For instance, the solar wind input to the magnetosphere can be assessed via solar wind plasma and interplanetary magnetic field observations that have been processed to account for propagation from the spacecraft to the nose of the bow shock. Also, the events that THEMIS observations can be put in spatial and temporal context by using magnetic indices and observations from other spacecraft located within the magnetosphere and from ground-based magnetometers, auroral imagers, radars, etc. In particular the full set of ground observations, from THEMIS and from non-THEMIS sources, is required to assess whether a storm or substorm is in progress and the timing of events. This is the type of a study for which a fully functioning VMO would be ideal. There are many

sources of data. The investigators need to determine which of the required data are available. The VMO should be able to lead the scientists to all of the required data: (magnetosphere) THEMIS, Cluster, Geotail, Synchronous orbit spacecraft: LANL, NOAA; (propagated solar wind) ACE, Wind; (Ground observations) Magnetic field data from the THEMIS array, and other magnetometer arrays; Auroral observations from THEMIS and non-THEMIS observatories, Radar, Indices.

After helping the investigators locate the data the VMO must help them access the data. The user will be able to access all of the data needed for the precise time interval wanted without having to query each data supplier's web page. With so many sources of data it is virtually guaranteed that they will be provided in a variety of file formats. This means that the VMO has a further responsibility to help the users get the data in formats that are easily read by the tools most frequently used in data analysis. The VMO will provide access to simple services such as plotting routines to help the users organize data.

It should be noted that much of the data needed for a study such as the March 1, 2008 case is not formally magnetospheric data. For instance, the study needs solar wind data, auroral data, and energetic particle data. It is the responsibility of the VMO to provide a single view into all of these data and to directly deliver the data from all of the sources to the user. To do this, it must be able to query the metadata registries of the virtual observatories that are responsible for these data and display the search results in a form useful to magnetospheric researchers for access and analysis. In short the user should be able to say I have found all the data I need, and the VMO will deliver it in the user's format.

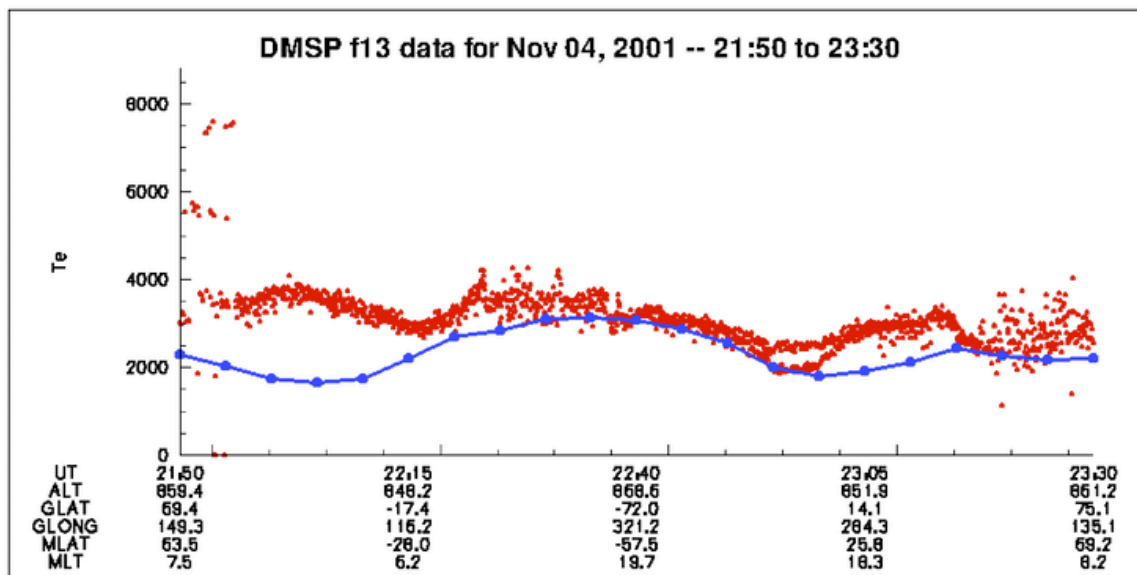


Fig. 4: Electron temperature from DMSP (red) and a magnetosphere model (blue).

A Model-Data comparison use case. Figure 4 above shows a test of a CCMC magnetosphere model that uses a generated comparison between electron temperature measurements from the DMSP satellite (red dots) and output from a model (blue dots) for a specific day during a magnetic storm. The DMSP data was dynamically downloaded from the University of Texas, Dallas, while the model was housed at NASA-GSFC and was also run dynamically. The comparison was done by a user at the University of Michigan's Virtual Model Repository, which allows users to make such comparisons directly, downloading the data and remotely running the models as needed. This specific example shows that the model is capturing most of the general structure in the Southern hemisphere and up-leg of the Northern hemisphere pass, but at the beginning, the trends are vastly different. In the past such comparisons could take days of identifying data, downloading, reformatting, and generating plots. When these comparisons are easy, they become routine and allowing orders of magnitude more data-model comparisons and thus much faster and better model development.

Collaborations. Science benefits from collaborations involving the exchange of ideas, theories and information. Collaborations are enhanced in many simple yet powerful ways by the HDMC. Some examples are:

- I have an event list. Instead of posting it on my personal web page, I submit it to HELM. Now both VxOs and researchers can refer to it with a URL and seamlessly link to it with a SPASE ID.
- The NOAA SEM2 data set has long been known to be a difficult data set to access because it has never been converted into a standard format. Because users were interested in applying correction algorithms to this data set, as a service to the community, ViRBO worked with scientists who had dealt with the data set to put it in a more accessible form, now available through ViRBO.
- I have a radiation belt data set. Instead of just posting it to my personal web page, I work with ViRBO to identify ways to efficiently communicate this data set to the rest of the community, even if it stays at my site.
- A scientist has a proton model that was developed as a part of a proposal. A deliverable was a stand-alone program that computes the near-Earth proton flux as a function of various parameters. The model developer works with ViRBO to create a service that outputs the flux given an input date.

Interdisciplinary Reality: Real overlaps and real problems

It is often said that a major reason for Virtual Observatories is cross-discipline research. As noted in the above, this is only one of many reasons, but it is worth considering what it means in more detail. The canonical cross-discipline example in Heliophysics is following space weather events from their initiation to their magnetic and energetic particle effects on the Earth and its denizens. It is useful to examine what the likely reality of such cross-discipline use would be. In practice, for Heliophysics research, people will be more interested in causes than effects. Thus, for example, solar physicists will rarely be interested in much beyond the corona, although images out to ~ 1 AU now provide increasing incentive for collaboration on in situ observational studies of the heliosphere. Heliospheric physicists are increasingly examining the sources of heliospheric transients and fluctuations. A relatively small fraction of solar and heliospheric researchers study the effects of solar phenomena on the Earth. Conversely, those who study phenomena closer to the Earth in the magnetosphere and below are most interested in the input functions from the Sun as measured by high-energy particle and photon fluxes, and the solar wind plasma and magnetic field parameters, rather than the details of solar phenomena and mechanisms. The ionosphere/thermosphere/mesosphere is largely unknown territory to most solar and heliospheric researchers. The strong coupling of ITM processes and the magnetosphere, which goes in both directions, has made these communities interact more strongly, with missions such as THEMIS explicitly bridging the gap. Modeling efforts increasingly involve coupling of models of these various regions including the radiation belts and ring currents. In general, phenomena farther from an event of interest tend to be reduced to indices or event lists. What was Dst? Was there a solar CME or a flare of relevance? What was VBs in the solar wind? In the short term, recognizing the realities of interests of one subfield in others will allow us to focus our efforts. As it becomes easier to approach other fields through improved tools, the gaps will become easier to bridge. Increased understanding between the different subdisciplines will lead to deeper insights, and, in turn, to the deeper and broader understanding of the global Heliospheric system that we hope to achieve.

Current Status

Inventory.

The original idea of the Virtual Space Physics Observatory (VSPO, <http://vspo.gsfc.nasa.gov>) was to provide access to all heliophysics products through a 3-D visualization application ("ViSBARD"). This is still a goal, and we have generalized it to include access from any application that can connect to Internet resources in the appropriate way, which includes such common programs as the Interactive Data Language (IDL). However, we realized early on that a more attainable goal was to have a web interface to all products, where the access could be through a link to service that rather than always being able via direct access. Essential to success was to know about and register as many products as possible, so we began to survey data products and to use an early version of SPASE to describe products in simple but useful ways, always with an associated access link. Using Web Service, ftp, and other access for SPDF, VSO, OMNI and other data sources, we have added more direct data access. The current list of products (now described using the latest SPASE version) includes nearly all the web accessible data from the NASA operating missions, all of the most popular products from CDAWeb (accounting for $\sim 95\%$ of all accesses), a wide variety of very useful browse products including movies of the Sun from Yohkoh and SOHO and overview plots from most of the active missions, a large collection of legacy datasets, and many non-NASA data products. There

are also pointers to models such as at the CCMC, with an interface to allow searches for model runs, and other resources such as a tool to answer, “Where was that spacecraft and what else was up then?” The VSPO was very well received in the previous Senior Review of Data Centers and became part of the SPDF at that time, thus making it the natural part of the infrastructure to be the face of the Heliophysics Inventory. The usage of VSPO has grown steadily, based on a constant measure of Web accesses that exclude web crawler and GSFC/APL hits. The current level of ~23,000 accesses per month is roughly twice what it was a few months ago. Currently we are working on adding all CDAWeb and VSO products to the Inventory, and on replacing simple VSPO SPASE descriptions with the much better ones that are VxO or provider generated. VSPO is available as a Web Service.

Virtual Observatories.

After the VSO, the discipline specific VxOs were formed by individual research proposals in response to a standard NASA Research Announcement. The idea was to let each subfield decide what would work best for that area. The initial VxOs were selected in response to a 2006 NRA for three year funding starting in FY 2007, and thus their funding is nearing its end. They will be continued, subject to the findings of this SR, under the HDMC as infrastructure projects, just as VSO and VSPO are now funded in the Data Center budgets. The initial VxO proposals covered the standard range of subdisciplines divided largely by region: Heliosphere (VHO), Magnetosphere (two, VMO-Goddard and VMO-UCLA), Ionosphere-Thermosphere-Mesosphere (VITMO), and Radiation Belts (ViRBO). Not all Heliophysics data products seemed adequately covered in the work proposed by these groups, and this led to the subsequent selection of VOs for “Energetic Particles” (mostly in the heliosphere; VEPO, with an FY 2008 start) and “Waves” (largely radio and plasma waves; VWO with a 2009 start). Finally, it has been clear all along that integration of observations and model output is essential for understanding, so a “Virtual Model Repository” was selected with an FY 2009 start. The VHO and the Goddard-based VMO decided from the start to share “middleware”: the software that links users to repositories in a uniform way. The two VMOs proposed complementary tasks and agreed to combine efforts right after selection. Both VEPO and VWO have also decided to share the VHO infrastructure in the interest of economy. In FY10, “SuperMAG,” an NSF-funded project that unifies access to ground-based magnetometer data, will be funded by HDMC as a VxO-like data provider, and they will add access to auroral imagery from spacecraft as a complement to their extensive magnetometer access. (Their HDMC start was delayed a year, pending the launch of their basic services.) A summary of the VxOs and other data sources, with web links, can be found at http://lwsde.gsfc.nasa.gov/hpde_data_access.html.

Each of the VxOs had as its charter to provide unified access to a significant set of data products within its domain. The teams combined discipline specific science expertise, software developers, and representatives of various relevant missions and repositories. In most cases, some funding went directly from the VxO to the mission data providers. The two major tasks were to describe the varied data products in a uniform way, and to design software to deliver these products to users. The initial proposals predated any formal Data Model, and thus neither this nor any other standards were required. In time all the VxOs have agreed to abide by some standards of interoperability, including at least basic support of the “SPASE Data Model” (see below) to describe the data products they deal with. As with any grass-roots efforts that are undergoing a subsequent integration, there are a number of issues of translation between metadata that have varying degrees of difficulty. The VxOs have taken quite different initial routes to serving data, with one group providing an integrated interface and a structured search route to the data that includes browse views (VITMO), and many of the others initially providing more basic data listings and menu choices with the option to plot datasets thus retrieved. Despite the various obstacles, the VxOs now provide access to a substantial fraction of Heliophysics data. The lists of products covered and planned are outlined in the VxO specific sections below. The issue of uniform delivery of resources is being addressed, and this, along with the issue of completeness, will be a major focus of the work as we proceed to the next stage.

As another grass-roots effort, the SPASE Data Model itself has taken a longer-than-expected time to mature. Since the most “natural” description of a dataset depends both on the nature of the data and on the reason for describing it, no one Data Model will perfectly fit all cases. The process of making tradeoffs resulted in many long discussions, but we now have a Data Model version 2.0 that is stable and being used extensively. *As a rough quantitative measure of where we stand in terms of the completeness of the application of the SPASE data model, based on our Inventory and other knowledge, we estimate that about 25% of the space-based (primarily NASA) data products have been described to date.* Some VxOs, such as ViRBO, have a smaller set of resources to deal with, and thus they will have completed the description of a larger fraction of their resources within their initial grant period. The VMO challenge is much larger, and this was the main reason for awarding two VMO grants. The two VMOs act as one so that there is no duplication of effort.

On a final note in this section, the original concept of a virtual observatory was as a “small-box” in which the Virtual Observatory functions as integrator of access to resources from multiple data providers who store their own data; the user obtains the data directly from the provider using the VxO provided pointers. This model works well in fields where there pre-exists a level of uniformity in formats and tools, such as in Solar Physics, that permits nearly transparent exchange of data resources. This is not the case in the broader Heliophysics Data Environment. While it is possible to build quick ad-hoc solutions, these often lead to less uniformity and they do not age gracefully because of the overhead related to adapting to each new addition. To achieve broader integration and coordination, including with international partners, the HDMC has been active in establishing a uniform data environment through its support of a common data model (SPASE) and an expanded role for Virtual Observatories. Thus, while the VxOs generate the metadata to describe resources, they also have at times helped to reformat data to aid researchers and acted as (temporary) repositories for “homeless” data. They monitor their subfields to assure resources remain available, and create tools for use throughout the HPDE. The HDMC coordinates the activities of the VxOs and the tool development activities to reduce overall costs by minimizing overlap and enabling the sharing of development efforts.

Resident Archives.

The initial Resident Archives were formed informally as extensions to mission funding, and thus Yohkoh data continued to be served by Montana State University (<http://solar.physics.montana.edu/ylegacy/>), and SAMPEX data came to be served by the Ace Science Center (<http://www.srl.caltech.edu/sampex/DataCenter/data.html>). After the initial success of these experiments, the idea of Resident Archives was formalized in the Data Policy, and the first round of “official” RAs is being funded this fiscal year. The main missions of relevance are Polar, Yohkoh, TRACE, SNOE, and SOHO, and each of these is starting their respective archives. In the case of the Polar mission, more than one archive will be established, with the electric and magnetic fields instruments in one site, one particle site (TIMAS) in another and imaging in another; not all the Polar datasets are currently slated to be served by an RA, but there are plans to complete this process.

A variety of models for the RA are being used, including one site for the whole mission (Yohkoh); multiple sites for one mission (Polar); and multi-mission sites (ACE and SAMPEX; SNOE and Polar TIMAS). In some cases, data from one mission will become part of a subsequent mission’s archive (some of SOHO served by SDO; some WIND data served with STEREO data). The plan is to accommodate what makes sense in each case, due to groupings of scientific expertise, shared data systems, or the natural connection of missions. We expect that there will be consolidation of various sorts as datasets age. The serving of data through VOs means that the physical location of the data does not generally matter, although some services more naturally resided at the archive site. For example, TIMAS provides the user the option of producing various cuts or integrations over an underlying particle distribution, and this service uses software tailored to this purpose at the TIMAS site. In most such cases, eventually a set of products will be produced by running the local software with common requests to prepare data for a Final Archive and, in fact, a number of TIMAS data products are already available at SPDF. In this context, a number of missions (e.g., THEMIS, STEREO, and C/NOFS) are using CDF and producing many data products as files rather than via software, and in these cases a Final Archive is simple: the final datasets produced by the mission are the Final Archive products, and they are easily served. Even with data stored elsewhere, Resident Archives are still appropriate to provide scientific expertise and possibly additional ongoing mission services.

Data Upgrades.

High quality data are the core of the Data Environment, and are what makes scientific progress possible. In the past, the provision of legacy products of immediate utility to the community was not always a priority, leaving a number of very useful datasets inaccessible. There have been a number of routes over the years to restoring or improving data products from former missions. The HDMC has formalized the process for significant Data Upgrades (more than what, say, SPDF could do as a small task) to be that a short proposal is submitted in response to the same ROSES NRA call that is used to initiate other HDMC components. There have been many significant datasets that have become available through this route. The proposal call asks that these be delivered in a form easily available to VxOs, and this has generally been possible, although some of the Upgrades were more focused or have mainly concerned the replacement of aging hardware at modest cost. A complete set of the Upgrades that have been and will be performed be found by examining the complete sets of approved proposals the on the HPDE website; these will be presented more directly in an upgrade of the site. The Upgrades include unique, high-resolution data from

the two Helios spacecraft in the inner heliosphere, data from the highly productive Dynamics Explorer spacecraft; hardware upgrades for Ulysses HI-SCALE, LISIRD (solar irradiance data), and ACE ULEIS; a high-resolution and expanded version of the very popular OMNI dataset of conditions upstream of the Earth from many spacecraft; additions to CME catalogues; and improvements to data access for ACE, SWRI-held and other datasets. More such Upgrades are currently in the initial funding stages, and we expect this part of the program to remain vital to overall success.

Value-Added Services.

As the VxOs began to offer the promise of delivering significant amounts of data, the program added a services component to the Research Announcement. In the first year, with FY08 starts, two proposals were funded in this category. One was to generalize software tools developed for the RHESSI mission to be useful for other solar data, linked clearly to VSO. This project has made excellent progress, and it will make it possible to obtain and use wide variety of solar datasets. The other project was to do initial work on developing a “semantic middle layer” that will enable other software to understand the contents of each of the data products in a mission-independent way, thus allowing scientists to focus on using the data rather than reading it and making it useable. This pilot project will be incorporated into the proposed work below.

The next, most recent, additions to the value added services include an initial study of inversion algorithms to obtain appropriate physical parameters from Energetic Neutral Atom images and a software package that will facilitate the interpretation of ground-based magnetometer data in the context of VMO. A third project is developing a visual interface to solar data that will allow the user to find and see browse images of events and features on overlay images of the Sun, and to make movies of data for selected regions and cadences from a wide variety of solar instruments. This will provide a new and powerful route to browsing and obtaining solar data. Finally, the Heliophysics Event List Manager project will provide a means to unite event lists that are stored in many formats, and to perform intersections and unions of the resulting time intervals. Currently, users often must examine separate catalogues and systems to make lists that must be combined manually. There is often considerable knowledge to be gained simply from knowing that, for example, a particular type of solar event is correlated with specific interplanetary or magnetospheric responses, and thus the event lists can be used for science by themselves, as well as part of searches, as is done already in VITMO and VSO. The Event Manager will enable VxOs to use a wider set of events, thus increasing their utility.

Proposed Work

A subsequent section will discuss the specific VxO plans. This section will present the proposed work in terms of the basic problems that we face in the Data Environment. We plan to:

- (1) Create an inventory of all resources, assigning each a unique ID (Part of Task 1 on p. 3)
- (2) Register the resources and provide means for updating and using the registry (Tasks 1, 2)
- (3) Provide unified finding of and access to the resources, by subfield and generally (Tasks 2, 7)
- (4) Allow easy browsing of all data (Task 3)
- (5) Provide user desired data in desired formats (Task 4)
- (6) Develop new tools and services (NRA; Task 5)
- (7) Describe all resources in a uniform way (with SPASE), and well-enough to be used (Tasks 6, 7)
- (8) Maintain and increase data quantity and quality (Tasks 8, 9)

(1) Inventory and IDs

The first requirement for an effective data environment is that high-quality data exist and are well described at the appropriate archive. The second requirement is that all data products are inventoried so that scientists need not “Wonder if they have all the data available” (Fig. 3). It will also be very valuable to assign a unique ID to each data product in the environment. Once we have such a set, then many things become simpler. For example, citations to data can be of the form “time range ... from dataset ID = ...,” thus allowing journal articles to refer to data as published items, just as we do with other publications now. The journals will reasonably require, as, for example, the Journal of Geophysics Research does now, that such IDs be stable over time and held by a long-lived entity. The SPASE IDs will fill this role for Heliophysics (see also below), since they will be independent of access URLs and repositories. Data systems will also be able to refer to data by time range and ID, possibly along with other parameters, and this will enable a variety of services including the ability to obtain direct access to datasets from

analysis programs such as IDL or perhaps even from search engines like Google. Browse products such as GIF images of graphs or JPEGs of images will be easily referenced by their IDs and associations between browse and definitive data products will be natural through the association of IDs.

We now have a preliminary but fairly complete listing of data products and other resources, such as observatories and instruments, that will need to be listed in the Heliophysics Resource Inventory; a large fraction of these are available through VSPO. (Go to <http://vspo.gsfc.nasa.gov> and choose, for example, the Observatory list at the left or the “View Current List” on the right.) ***In the next year, we will be completing a Data Product Inventory based on all available information on what datasets do or should exist. It will be kept current.*** It will include the VxO or Data Center that will be responsible for maintaining the metadata for the resource, the type of current archive (Mission, RA, Final, Deep, or Other), and the unique ID, along with the minimum of information needed to identify the resource and its content and format. Most of the metadata needed will be in the SPASE descriptions, and we will generate an online viewable/sortable version of the Inventory based solely on this. (VSPO does a simple version of this now.) For archival purposes, we will extend the basic database to include deep archive sets and some that are expected from missions but do not yet exist. We will also associate a brief qualitative assessment of the product status (documented; only level zero; needs recalibration, reliable final version, etc.) with each product. It should be noted that inventories for Heliophysics data have not lasted in the past, and that the current NSSDC Master Catalogue, while very useful and part of our efforts, is not complete. The difference this time is that there are more compelling reasons for the Inventory, namely the desire and the technical capability to obtain access to a full range of resources, and there is a formal Data Policy, endorsed by the community, stating that the Inventory is needed. This was agreed upon in the earliest stages of the Space Physics Data System, and has been implemented in other areas at NASA and elsewhere. We will take advantage of the many improvements in access and software systems to provide simpler and more direct access to resources than typically provided by current inventory systems.

(2) Registry Services and Access

Resource IDs and associated SPASE descriptions are most useful if they are organized into registries that are easily searched, browsed, and used for data access. Each registry of metadata needs be easily updated and harvested. For a small set of products, a text editor used with a registry consisting of a simple set of files could be an adequate, but more typically a VxO or Data Center will maintain a database, either with commercial software such as Oracle or using open software such as eXist. The latter has the advantages of being free and naturally compatible with SPASE because it is XML-based, but historical or system design reasons may preclude its use. It is very helpful to have tools that allow additions, modification, deletion, and harvesting of these more complex databases. We have the components of a system to manage our registries, and will connect them into a complete system within a few months.

A core component of the HDMC registries is the Git version control system that allows us to maintain HDMC-wide common registries of observatories, instruments, repositories, and people (the Space Metadata Working Group, or SMWG, set of resources). Updates require human checking, as they always will, but otherwise harvesting is straightforward, and the metadata entries are easily accessible. VMO, VHO, ViRBO, VWO, and VEPO all use this system for their SPASE descriptions and version control, and we plan to include the other VxOs and VSPO in this system. Our registry tools will use Git but will be designed to allow flexibility for non-Git users and to be ready for possible needed upgrades. Git is a widely used, efficient, open-source tool with a large user base that includes the development of the Linux Kernel (see <http://git-scm.com/>). In addition to the standard harvesting tools, we have built basic editing and submission/update tools that work with Git. These have started in particular VxOs, but will be made into services that all can use, exposed through the SPASE website. ***In the next six months we will finalize a registry architecture based on the above ideas.***

For a general registry, VSPO is harvesting all of the SPASE records from VxOs and Data Centers (e.g., SPDF, host of VSPO) and maintaining a current, complete list, storing them and serving them from an XML database. Others may do this as well, but we want to have at least one place where anything available is easily accessible. ***The initial SPASE descriptions created and used by VSPO will be replaced over the next year or two by more complete descriptions provided by VxOs and others; this process will generate the final IDs for a comprehensive set of resources and provide links to VxO services.*** VSPO (to become the “Heliophysics Resource Gateway”) provides searches based on free keywords (Google-like) or based on SPASE elements, chosen in any order. This provides an easy route to finding resources. Once found, VSPO provides the SPASE metadata and allows the user to obtain further descriptions using Information URLs or the data itself through Access URLs. It provides links to relevant

VxOs and other services with greater functionality, and also uses direct links to VxOs and Data Centers to deliver data using their services while staying in the VSPO interface. Thus, VSPO is the “face of the inventory” in a way that far exceeds simple lookup. It is like an electronic card catalog for a library that also delivers books and CDs to the user’s computer. We believe that this increased utility will help considerably to maintain community interest in a Heliophysics Resource Inventory, and thus to ensure that this Inventory will not suffer the fate of some before it.

(3) Finding and Accessing Resources

The idea of a VxO is to build on a backbone of registered resources to provide uniform access to a set of products and services that are useful for community “x”. We now have nine VxOs within NASA, and various other similar activities, including VSTO (originally NSF funded), the European “HELIO” project that is parallel to HDMC, and non-NASA “VOs” such as GAIA and SuperMAG that will provide uniform access to some subsets of data. It does not make sense to have a dozen places for “one-stop shopping,” and this is not what we intend. Many of the current activities are developing essentially similar interfaces, so we plan to consolidate over time. The indispensable virtue of a VxO is the ability to understand the data in its domain so that it can provide the required metadata to make the resources in its domain comprehensible to users and easily available. For this role, the current set of VxOs provides good coverage of all of Heliophysics. This does not imply, however, that each of these groups needs a separate interface with links to the others. There are five VxOs—the two VMOs, VHO, VEPO, and VWO—that are sharing the same middleware; some of these will share the same database and interface. VEPO is mainly concerned with energetic particles in the heliosphere, and these cannot be understood outside the context of the heliosphere in general, and VSPO is already using VHO for its products. The two VMOs are working as one in covering products, and their initial experiments in interface design will lead to a single interface. Products concerning waves (largely radio and plasma waves) described by VWO will be relevant to VHO, VMO, or ViRBO depending on the particular product, and these products are not nearly as useful outside of the larger context. Here again, consolidation will be important. Ideally we will have a set of resources and tools that can be combined by users as they see fit, and this is our ultimate goal. In the meantime, we will strive to simplify the user experience.

The primary function, then, of the VxOs in the next two to three years will be to complete resource descriptions and easy access to the data within their domains. Some of the latter task will be separated out into activities that clearly cross domains, and thus will be funded separately such as the visualization and communication tasks described below.

(4) Browsing and Visualization

An extremely useful way to find new phenomena and understand data is by scanning through large numbers of plots by eye. This is frequently an efficient way to do searches, especially in the early stages of an investigation, and it is a very easy way to understand the context of data. For example, examining a quick plot of OMNI data will usually provide the information on the solar wind conditions driving a particular magnetospheric event. Thus browsing and visualization are essential to the easy use of data. On the other hand, when it comes time to make finished plots for publication, each user and journal will have different preferences, and it does not make sense to produce yet another industrial strength plotting package. This is too much of a task for HDMC resources in any case.

There are a number of useful approaches to browsing data. One of the simplest, used extensively by VSPO via both separate services and a more unified “DataShop” interface, is to take advantage of the GIF images of plots or JPEG/MPEG pictures and movies made available by data providers. There are very helpful overview plots from most of the major current missions, as well as movies from a number of the solar instrument sites. These “Display Data” (in SPASE terms) are available through various web or FTP sites, and in most cases DataShop can use regular expressions to parse the locations of images for specified time ranges. Services often allow simple scrolling in time with “next” and “previous” buttons, and the DataShop approach allows recalling many images at a time for easy comparison. We expect that as these overview sites become more widely used, they will be adopted as easily added services by more providers. ***We will include easy access to browse plots from others and made by the VxOs themselves to make it easy to understand the content and relevance of any dataset to a particular science investigation.***

A different approach is to make plots on the fly based on the underlying numerical data product. This can be done by a service at the provider site, as with CDAWeb, the ACE Science Center, and many other sites. However, this will not cover all products, so it is useful to have an application that will read all but the most idiosyncratic data formats. The application we are using for this, called “Autoplot,” originated in the ViRBO team, and it has been

adopted by many other VxOs. Autoplot is intended to perform the extremely useful task of automatically producing a sensible preview plot given data from many different data sources on many different platforms in many different user modes (browser applet, desktop client, and in a server mode). ***We will continue to develop the Autoplot application to make it robust and efficient for VxO and independent use.*** Autoplot can be seen in action at the ViRBO site, and is documented at <http://autoplot.org>. While it is a much less urgent task, we have linked VSPO (in a prototype) to a 3-D visualization tool, called ViSBARD (<http://spdf.gsfc.nasa.gov/research/visualization/visbard/>) to provide 3-D browsing and analysis of scalar and vector time series data. It allows the user to see the realistic context of observations as well as making it possible to understand dozens of time series at once. ***Although as a lower priority, we will continue to work on the integration of tools such as this that will give users a greater ability to understand multi-instrument and multi-mission data within the VxO environment.***

(5) Solving the “Formats Problem” and “Dataset Runs-on-Request”

After the provision of high-quality data and making it easily found and downloaded, the next problem facing a scientist is again well summarized in Fig. 3: They don’t want to “Spend time and money getting data into a useable format.” We propose to solve this problem. The problem has decreased considerably in the last five years or so, since many missions have adopted a standard format for their data. New missions in Heliophysics are mainly contemplating the use of FITS, CDF, or HDF for their data. This is a great improvement over having level-zero files in EBCDIC, VAX binary, or some custom encoding as the primary mission products. Nonetheless, there are still a number of formats to deal with, and ASCII files, while they can be read by nearly any application, are not self-documenting. ***We will address this problem by providing a “universal reader” (within some decreasingly important limits) that can produce a variety of common output formats, using a common representation internally in the software. The common middle layer can eventually be tapped by applications, thus avoiding extra layers of translation.***

Longer term, the solution to the formats problem is, simply, to use at most a few of them. The reason we now need the software in the middle is the large quantity of legacy data, and the need to deal with ASCII. We also cannot control non-NASA formats, so this problem is likely to persist at some level. The Space Physics Final Archives would have an easier job if everything were put into CDF for space physics data for easy serving, as solar images are now in FITS. Some fields, especially as they get closer to Earth Science, may still want HDF/NetCDF. The convergence to a few formats is happening de facto in the community, but discussion with the HPDCWG made it clear that the community is not quite ready to see this as a NASA requirement. The Data Policy does have a recommended set of formats (CDF, FITS, HDF, and “ASCII”), but not as a firm requirement. The implication is that use of anything else needs to be clearly justified.

Part of the Autoplot application is intended for the above purpose since provides easy means to understand the syntax of many file types, and provides basic output files. Another application we are developing is “DataShop,” which provides uniform access and output of many different HP datasets. Internally, these two tools share an encouragingly similar solution to the formats problem in which native data content is mapped to an internal representation that is independent of format. This many-to-one conversion on data input is the most efficient way to provide data format conversions. ***We propose to take these separate efforts for data handling and integrate them into a single project focused on a community-endorsed mechanism for handling data of different formats.*** The main goal of this project will be to create a universal reader library consisting of easily generated XML files for each product that will be able to read any HP dataset (with the possible few exceptions for oddly formatted outliers) and provide output of that dataset in a wide variety of formats. The use of a modularized design, as is already done in both Autoplot and DataShop, will make it possible to easily and independently add new input and output types.

One aspect that sets this project apart from the others is the amount of time and effort spent dealing with non-plotting issues such as making sure the code runs on multiple platforms, is easily extendible, is responsive, has an intuitive UI, has bugs fixed quickly, etc. The development is open source, and there have already been contributions from outside the core development team. While it is being developed as a “browse” application, as discussed in the next section, in this context it is the fact that it can read a wide variety of formats that is essential. The desktop client version of Autoplot will provide a way for a user to read nearly all the files that are likely to be encountered in our current data environment, and thus users will not have the face the formats problem again.

We will use the format solution to develop a Dataset Runs-on-Request that will be a powerful way to “Create versatile datasets” as requested by the scientist in Fig. 3. We have also developed within the HDMC a number of

other tools that, along with those in Autoplot, will enable a service for larger data requests. We call these “Dataset Runs-on-Request,” analogous to the CCMC Runs-on-Request for models. The VHO/VMO groups are running parallel code to produce uniform, modest resolution datasets with statistics for use in searches. The SPDF has at various times contemplated a “batch processing” facility for large requests. The ViSBARD, a 3-D visualization application mentioned above, has within it the ability to merge datasets from different instruments, possibly on different spacecraft, with a chosen time basis. ***We propose here to use our collective expertise to provide a tool that would allow subsetting by variables and time, merging data across file boundaries, merging data from different datasets, choosing either the cadence of a particular variable or a fixed cadence for the result, and choosing the granularity of the resulting files in terms of size or time-range.*** For a particular format, such as CDF, this is not a difficult task, and our current tools just need to be put into a callable service. Data would be interpolated or averaged as needed or desired, and a missing data flag would be chosen or interpolation could be used. Output formats would include MatLab “mat” files, IDL savesets, CDF files, and ASCII in the desired form. Such a service would be of great benefit for many studies. For example, the Ulysses plasma data varies between 4- and 8-minute resolution, whereas the magnetic field is as much higher cadence. Many studies require both datasets, and for correlations, calculation of plasma beta, and many other purposes, a combined, uniform-cadence dataset would be useful. In fact, such a set has been produced a number of times over; we propose to allow the user to make whatever sets are needed, as needed. It will likely be useful to document and archive these for other users, but the tradeoff between managing that archive and allowing users to simply run jobs again needs to be evaluated by experience. There will be a number of such datasets that are likely to be of wide utility, and we would produce some of these as we determine through interactions with the community that they would be useful. In all cases, provenance information would be clearly documented. There are already examples of this sort of dataset preparation in the “OMNI” and “COHOWeb” summary combined magnetic field and plasma datasets produced by J. King; these have not produced any negative community reactions because they are carefully documented.

(6) Tool and Service Development

The Data Environment we envision here allows individual researchers to develop tools and services that can take advantage of the uniformity of description and access. This is an open area for new initiatives, and it will be important to foster good ideas. The funding in these cases is and would be dependent on having a very clear plan for transitioning the tool or service to being a supportable part of the infrastructure as part of a VxO or Data Center. Examples of possible projects include the improvement of 3-D visualization tools (such as ViSBARD), and the development of data mining tools and services. Very frequently researchers are faced with the questions: “Are there any other events like this?” or “What characteristics do events like this share?” There are now automated methods for finding the answers to such questions, some in use in Heliophysics, but they continue to be difficult to use, mainly due to the need to prepare the data properly. These suggestions are only that, and the directions taken in this area will depend on the ingenuity of the proposers. ***We propose to continue to fund Value Added Services that will be tightly coupled to the architecture of the data environment.***

(7) SPASE Descriptions

A key to having easy searches across a wide array of data products is to have all the products described with common keywords; we refer to this set of terms and their relationships as a “Data Model” (see Fig. 4), although the term is used in other ways. Creating an adequate uniform language is no small task. Datasets differ in both large and subtle ways, and each person coming to the problem of categorization will start from a different set of assumptions. This has made the generation of an adequate Data Model very time consuming and at times contentious, and this, in turn, has slowed the writing of uniform resource descriptors. ***The international, interagency SPASE (Space Physics Archive Search and Extract) group performing this work, which involves all the VxOs and many Data Centers, has finally achieved sufficient agreement to arrive at “SPASE 2.0.0” that we have agreed will be the standard for the foreseeable future.*** The details about this model can be found on the SPASE website at <http://www.spase-group.org>. Any changes to this model in the foreseeable future will be in the higher decimal places, and can include, for example, backward-compatible additions of terms within controlled lists to accommodate new instruments.

The Data Model has been developed through several years of interaction of the open-membership SPASE working group, mainly through biweekly teleconferences and email exchanges. The working group has several goals:

- Facilitating data search and retrieval across the Space and Solar Physics data environment;
- Defining and maintaining a standard data model for Space and Solar Physics interoperability;

- Demonstrating the Model's viability;
- Providing tools and services to assist SPASE users; and
- Working with other groups for other heliophysics data management and services coordination as needed.

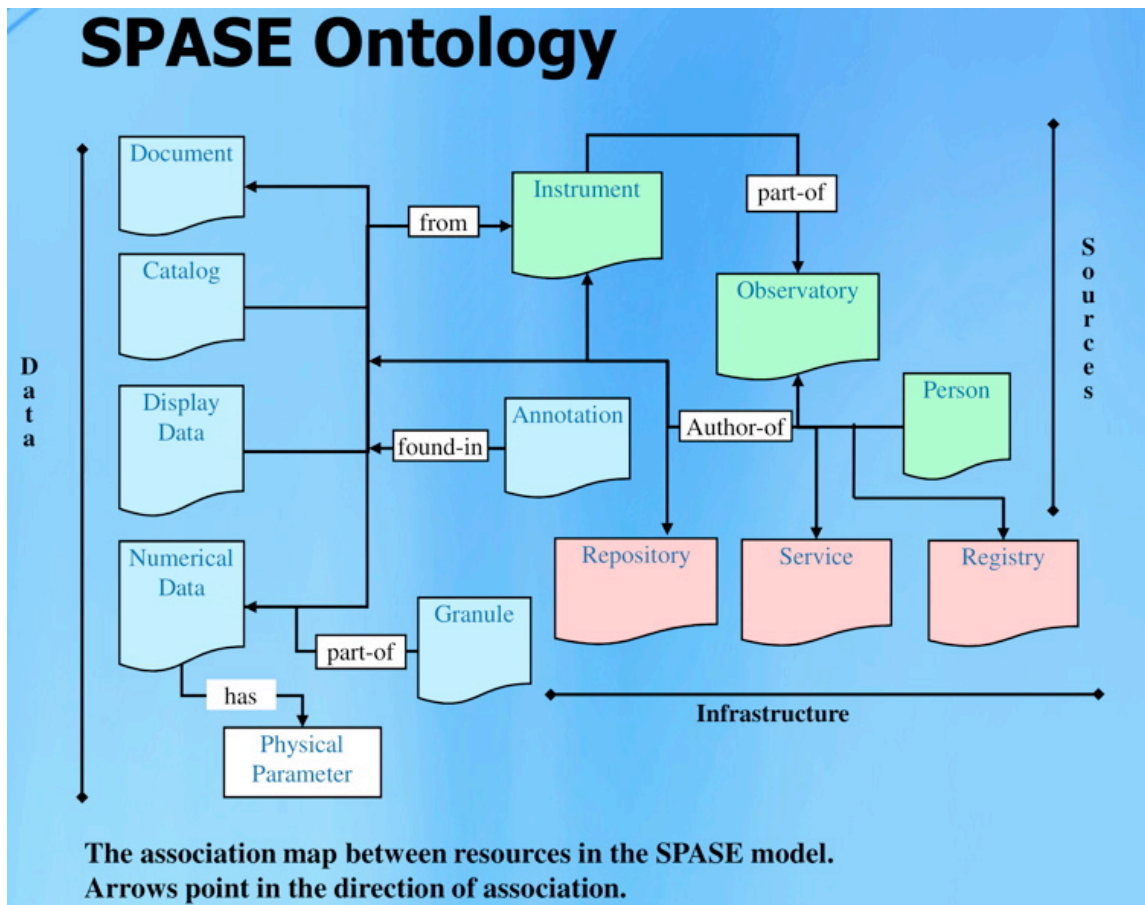


Fig. 4. The SPASE Data Model viewed as an “Ontology,” i.e., with indications of the relationships between the terms. Each colored box represents a “Resource Type” that has its own set of terms and structure. Many “header” items (name, description, etc.) are common to all resources, and Resources such as “Person” are very simple.

The SPASE Working Group is currently the only international group attempting to achieve interoperable global data management for Solar and Space Physics, and we will be coordinating with the HELIO project in Europe as it begins to tackle the same Data Model and other problems from a European perspective.

There has been considerable concern expressed in the community, for example through the HPDCWG, that SPASE will be too difficult and too mutable to be helpful, and that it will impose unfunded mandates on missions. The current Data Policy makes clear that highly detailed SPASE descriptions (with all parameters spelled out) are not required, and that any significant data description tasks will be carried out by VxOs or with their help and/or funding. We now have a clear plan for the completion of the SPASE description of a comprehensive set of Heliophysics resources, and funding for this plan is included in the budget for this proposal. The descriptions of existing resources should be complete within two years, and all the resulting descriptions will be harvested for the general Inventory so that we have a unique set of supported descriptions with which to proceed. Some VxOs will have SPASE descriptions more deeply embedded in their services, but this sort of difference will be invisible to end users. We are committed to stability in the SPASE Data Model, with changes limited to necessary additions to controlled lists and other minor changes that will not invalidate current descriptions. Any more significant changes will be done only with community buy-in and clear advantages for the expected science return from the data environment, and such changes are not expected during the completion of the description of the Inventory in the

next two years. *Now that we have a stable Data Model, progress can be much faster, and we propose to complete a comprehensive set of SPASE descriptions for Heliophysics data and other resources within two years.*

All the VxOs will provide web-browser access to data, but equally important will be machine-to-machine access that allows other VxOs, applications such as IDL, and services such as HELIO to take advantage of the unification provided by each VxO. This can be accomplished on an ad hoc basis, with each system defining its methods, but there are clear advantages to providing a uniform set of methods for asking for lists and descriptions of data and other resources, and for the data themselves. The approach we are taking is to use SPASE terms as keywords to describe a search with a simple XML file. We will also develop a “REST-ful” service in which the request is formed as a URL. The requirement placed on VxOs is that they be able to translate the SPASE QL query into the language it uses internally, which could be SQL, xQuery, or something else depending on the implementation. Since all the VxOs will be producing SPASE descriptions, and thus will know how their products are described in these terms, we expect that in most cases the required translations will not be difficult. All VxOs will develop a basic SPASE QL capability that allows simple harvesting of metadata and files; some will have the ability to allow complex queries with multiple restrictions based on a wide range of SPASE terms. Each service will advertise its level, and we anticipate that the levels will increase over time. *The initial implementation of SPASE QL is being completed by VHO and VMO, and testing by other groups is starting. Within one year basic functionality will be established and within two years the implementation will be complete, although improvements will be ongoing.* Examples of SPASE QL can be found at <http://vho.nasa.gov> by clicking on the “SPASE QL” button at the top.

(8) Data Quantity and Quality

As mentioned above, high-quality data are what the Heliophysics Data Environment is intended to provide, and what makes the science possible. Thus two aspects of the HDMC are nearly self-evident in their importance, namely, the Data Upgrades and Resident Archives. Due to many factors, there are still a number of useful but currently inaccessible legacy datasets that could be recovered for modest cost. Other data needs some further processing to become highly useful. For example, auroral images require careful background subtractions and other processing to reveal the activity of interest, and this was not always possible to do on a routine basis during a mission. Recent algorithm development will often enable better processing, but funding is still needed to carry out the Data Upgrade. Often a modest hardware investment will maintain or greatly enhance the utility of a data collection. *Thus we propose to continue the Data Upgrade part of the NRA call. We would continue to offer up to \$50K for one year for these projects, with the expectation that many grants would be smaller and that a few proposers would argue successfully that more resources are needed.*

The Resident Archives provide a means for mission teams to continue to serve and provide expert help with data that are still in use, and thus we propose to continue to set up RAs as needed. The initial set of these is just starting up, in terms of formal funding, but since they generally are just continuing current operations, the main issue is to provide funding to allow the operations to continue. As discussed above, there have been a number of approaches to RAs, and we expect many models to be successful. The next stage of this program will require that we formalize a light oversight capability. The HDMC, in cooperation with SDAC and SPDF, needs to assure that the RAs are “trustworthy archives” in the sense defined in Appendix F of the Data Policy, namely that they have proper backups, use checksums, maintain a disaster recovery plan, etc. *The Data Policy Appendix provides a checklist of requirements, and the annual reports from the RAs will be expected to show how these requirements are being met. We will ask for initial assessments from the RAs as they are set up, both to increase awareness on their part of possible issues and to help us with how best to deal with quality issues.* The goal will be to assist the RAs in preserving their data more than being “data police,” although *both the HDMC and Senior Reviews should ensure that the Archives are trustworthy as indicated in the Data Policy.* In some cases, it may be recommended to make copies of the data at a NASA Data Center or elsewhere to ensure safekeeping. Our experience has been that those who have produced and used data want very much to preserve it, so we do not expect difficulties in obtaining compliance with established best practices. The RAs will also be reviewed as part of the HDMC during Data Center Senior Reviews to determine if the need for the RA still exists, and if so what level of resources is needed for its task.

RAs will form as missions end. The decision to form an RA, as opposed to, for example, putting data products in Final Archives, is up to the mission team or others in the community interested in the project. As the data become less used, or the RA team decides, for whatever reason, to not continue to serve particular datasets, there will be a need to transition the data to another archive. In some cases, such as SOHO, the data may be served by a more

recent mission (SDO, in this case) that is similar and thus provides a natural home. In other cases, such as THEMIS, data products will already be in a Final Archive (in this case, SPDF), and thus already available. There may be needs in such cases to augment the Final Archive holdings or to serve the data in other ways as well, but this will be worked out on a case-by-case basis. The primary goal is to ensure that NASA missions, and, as possible, other missions will provide a legacy that will not require extensive Data Upgrade projects to keep meaningful data flowing. Other aspects of the Data Policy, such as the requirement of Mission Archive Plans, are complementary to the role of the RAs, and are intended to make the transitions between different phases of data archiving as seamless as possible.

Working Groups

We have set up four working groups that will focus on specific HDMC problems that need to be addressed across the Data Environment. The groups will implement solutions to some of the problems listed above in coordination with the larger HDMC group, and they will collaborate, as needed, with groups from other agencies (e.g., NOAA, NSF) and nations (e.g., HELIO). The working groups, each of which has a leader and four or five initial representative members, are:

- (1) Registry management services (resource description creation, harvesting, updating, editing, etc.);
- (2) Visualization (initial emphasis on Autoplot, but also including 2- and 3-D visualization; emphasis on browsing);
- (3) Data Processing Services (variously termed Dataset Runs on Request/Download Service/Time Series Service/Batch dataset processing--subset, merge, interpolate, translate, and average with multiple input and output formats); and
- (4) Interoperability (SPASE-QL, service APIs to link VxOs and other services to each other and to applications)

In addition, Event and Feature services will be organized around the approved value-added service Heliophysics Event List Manager project (uniform event lists, communication, event list logic, etc).

Roles of Specific HDMC Components: Individual VxO status and plans

VxOs all provide a number of functions. Central to these is discipline-specific expertise. The HPDE needs to have insight into each subfield that assures completeness of coverage and an understanding of what is available. The VxOs provide a means of assuring that data descriptions are complete, accurate, and useful. The VxO PI acts like an “embedded journalist” and attends (and possibly hosts) conferences and meetings to identify needs and ways in which their Virtual Observatory can meet their community’s needs given the resources of the greater HPDE. The missions are not generally going to have archival commitments to VxOs, although some, as with solar missions and VSO, may operate more closely. VxOs do potentially provide bidirectional pipelines of resources and interest between the mission data sources and the user communities. VxOs function as data environment focus groups for connecting resources to needs in the discipline communities, rather than just the providers of a set of software routines. The current VxO efforts have led to the improvement of a number of datasets and made them available.

All the VxOs are committed to provide a comprehensive set of SPASE descriptions of data products and the sure knowledge of where the registered resources are best found and acquired, and to provide a user-friendly interface for performing searches of varying complexity and acquiring the data thus found. They will also all implement accepting SPASE Query language requests at the level appropriate to their VxO. Finally, all VxOs are committed to working with their respective communities to maintain complete, up-to-date access to resources, and to identify and help to implement services to meet community needs.

The Original: VSO

One of the first suggestions that Heliophysics should emulate Astrophysics by providing one-stop access to a wide range of archival data products came through the over-arching proposal two Senior Reviews ago for a Virtual Solar Observatory. Based on the panel’s endorsement, a group of solar physics organizations established a “small box” of “middleware” to provide pointers to a comprehensive set of solar data through a single portal. The software provided a simple means to query a registry of providers and products for files that met a set of criteria, such as that

the data were for a particular time range and from a named instrument or were associated with a chosen data type or “nickname” (“H-alpha,” “Magnetogram,” etc.). The VSO then found and provided links to all files that met the criteria and the user could then download these, almost all in FITS format, to be analyzed by routines from the widely used “SolarSoft” library of IDL code or by other software. This approach, termed a small box because no data flow through or are manipulated by the VSO, has been successful, with a large fraction of existing space and ground-based solar data available, and with clear plans to add anything missing. Useful features have been added, such as a movie preview facility and the ability to use lists of events such as flares and coronal mass ejections to select search times. The SDAC proposal covers the other capabilities and plans for the VSO.

The VSO has provided a test case for other VxOs, showing that the basic ideas can work well. It is also a source for data for nonsolar groups. As such, and as part of the Heliophysics Inventory, the data products it serves need to be described with the same language as those in other subfields, and thus we need to create SPASE descriptions of what are now described using the VSO data model that predates SPASE. This activity has already begun, with a number of solar products of general interest served through the VSPO using SPASE descriptions. Many of the solar products are easy to describe because they do not contain a large collection of variables, as is the case with many other Heliophysics datasets. Thus we plan to rapidly produce the required descriptions through a VSPO/VSO collaboration. In a number of cases, VSPO directly accesses the solar data files using VSO Web Service software. To make the access to VSO data products more uniform, we plan to work with VSO to implement SPASE Query Language methods for data access. Working through the Heliophysics Event List Manager project (an HDMC value-added service project), we will also exploit and generalize the VSO event list capabilities.

Next set: VHO, VMO (U, G), ViRBO, VITMO

The Virtual Heliospheric Observatory has its roots in a project to unify data at L1 upstream of the Earth to better understand that significant region. Thus, initial efforts focused on providing access to magnetic field and plasma datasets from spacecraft in the heliosphere near the Earth. In addition, VHO has worked with VMO on providing parameter-value based searches and on SPASE QL. Currently, VHO is expanding its range of data descriptions and access, assisted by VEPO. The current data products include many from ACE, Wind, Helios, IMP-8, Genesis, Voyager, and Ulysses. The connections of VHO to its user community have facilitated the upgrading of access to a number of datasets. Previously unavailable SOHO CELIAS 30-second Proton Monitor data are now served from VHO by arrangement with the CELIAS team. The instrument team was funded by VHO to bring the data to publicly releasable condition and to provide SPASE descriptions of it. VHO also provided a home for MGS Solar Wind Proxy data that use MGS magnetometer data and a model of the Martian magnetic field to generate predicted solar wind pressure values at Mars. VHO worked with community members to provide Helios 1 and 2 combined Magnetic Field and Plasma data in a form unavailable elsewhere.

The VHO will be adding the rest of the data products (roughly half) from the missions they are already serving. In addition, STEREO and Messenger data products will be added. VHO will continue to work with VMO and the rest of HDMC to quickly implement SPASE QL. They will also include event-list based searches, complete Autoplot upgrades (e.g., add preset profiles for browsing), provide significant interface upgrades including a 3-D interface in later development years, and add services such as an IDL interface using the VHO API. The latter task should generalize to all VxOs via SPASE QL links, with help from routines included in Autoplot for file reading.

The Virtual Magnetospheric Observatory will provide access to a highly varied set of resources that cover a host of interacting, intrinsically complex regions. It has the largest collection of resources to be described of all the VxOs, and thus two separate groups were funded for this purpose. These groups now act as one, systematically dividing the tasks of resource description and developing complementary services. The VMO was instrumental in developing and implementing the Git-based approach that HDMC will use as a basis for Registry management. Along with VHO, they developed SPASE QL and parameter-based searching. The development of SPASE tools such as for schema validation and exploration of the Data Model was largely done by a VMO member who is also funded by SPASE. The VMO has made substantial progress in describing resources using SPASE, including data and related resources from AMPTE, Polar, Geotail, DE-1, IMP-8, Interball, Prognoz, and THEMIS spacecraft, in addition to a large collection of ground-based magnetometer stations and L1 datasets shifted to the Earth’s bowshock.

The planned activities for the VMO include finishing a comprehensive set of SPASE descriptions, improving the

web interface for easy data access, and developing basic access tools. The additional missions to be covered will include Cluster, Equator-S, FAST, Galileo, the GOES series, Hawkeye, IMAGE, ISEE 1 and 2, and more products from the current mission list. The collection of ground-based magnetometer descriptions will also be completed. The new web portal will combine text, keyword, and parameter-value based searches to improve search capabilities. The services to be added include a SPASE registry explorer (based on the SPASE Data Model explorer at <http://spase-group.org/registry/explorer>) to easily browse data products; a means to visualize the associations between related products; the ability to display “thumbnail” browse images of files in search results lists; a search capability based on the display of plots of data availability as a function of time; and the use of SPASE QL to save queries and pass them to other VxOs. Many of these capabilities will be of use to other VxOs. The VMO will make use of Autoplot for viewing and exploring the content of datasets, will actively participate in the development of a Dataset Runs on Request service, and will experiment with the use of the annotations of files to aid in searches and understanding. They also plan to provide event-based searches, develop initial event lists, and assist in development/management of event lists.

The study of the ITM region combines ground, atmospheric, and space-based measurements, with an accompanying array of types of data and file formats. The Virtual Ionosphere-Thermosphere Mesosphere Observatory has focused on locating and delivering large volumes of data, providing event, conjunction, parameter-range, and keyword based searches. A uniform interface allows all such restrictions to be applied, and additionally provides summary images of the search results to help further with the finding and understanding of event-related data. The results of searches are bundled and delivered to the user all at once along with descriptions and metadata to help with use. A unique feature is the ability to find coincidences between ground-based observations and remotely sensed space-based observations that generally provide information that is not at the spacecraft location. The datasets currently used by VITMO include all the TIMED data products, SuperDARN images and data, ACE upstream parameters, many geophysical indices, and the SNOE, Alouette, ISIS, ROCSAT, and DE data resident at SPDF.

Plans for VITMO include completing the addition of a comprehensive set of datasets, linking to other services, and upgrades of current services. The additional datasets include products from AIM, DMSP, UARS, and C/NOFS; and additional SPDF resident sets (Polar and IMAGE images, Aeros, San Marco, OGO-6, AE, OMNI), as well as the Digital Ionosonde database and TEC maps. The services that will be linked to VITMO include SuperMAG, VSTO, SSCWeb, and the CCMC. VITMO’s efforts are based on a data model that predated SPASE, but they will provide SPASE metadata and SPASE QL links to the unique ITM datasets they serve.

The Earth’s radiation belts are part of the inner magnetosphere, but they provide a unique set of problems of interest not only for scientific reasons, but also due to the hazards energetic particles pose for space-based assets. The Virtual Radiation Belt Observatory’s efforts have been unusual among the VxOs for requiring more “data upgrade” work due to a significant lack of public availability of relevant datasets. Thus, NOAA SEM-2 data was upgraded to be available in a generally useable form, and the GEO reanalysis of geosynchronous data to extend it to full orbits, requested by the GEM community, was completed and made available through ViRBO. By the end of the NRA grant period, data will be available with SPASE descriptions from many spacecraft, including Polar, SAMPEX, POES, LANL (various), HEO, GOES, CEASE, and OV. Related indices and OMNI data are also provided through the Web interface. ViRBO is providing surveys of many of its datasets very rapidly by caching uniform versions of the data locally and using Autoplot, which is software that they initiated and that others have subsequently adopted. They have also developed tools for Registry management. ViRBO is very active in community meetings to deeply understand the needs of the Radiation Belt community.

The plans of ViRBO include providing metadata and assistance with Data Upgrades for relevant data from TWINS, SAMPEX post-mission sets, full-resolution NOAA-14, THEMIS SSD, and AFRL’s DSX. Possible additions include data from DEMETER, Orsted, CHAMP, ROSAT, and TOPEX. They will work closely with RBSP to provide SPASE descriptions and related assistance. General services to be provided include an “L and L” (essentially, latitude based on included magnetic flux) service that is fast and integrated with SSCWeb; improvements in a data download service related to Dataset Runs on Demand; and upgrades to Autoplot for browsing of spectrograms. Possible long-range plans include support for assimilation models and a principal component analysis calculator.*

Newer: VEPO

The Virtual Energetic Particle Observatory (VEPO) is a Measurement-Type (a SPASE term) focus group operating in conjunction with VHO to support VxO discovery and access for energetic particle data products from heliospheric spacecraft. The energy domain of these products extends from GeV energies of galactic cosmic rays to suprathermal keV-MeV energies of seed particles for acceleration to higher energies by solar, heliospheric, and local interstellar processes. The VEPO group includes participating scientists who are, or have been, members of instrument teams providing energetic particle data from operational or legacy heliospheric missions including ACE, Helios, IMP-8, Pioneer, Ulysses, and Voyager. The team also provides expertise on ground level neutron monitor data as long-term records of heliospheric modulation for galactic cosmic rays and of ground level events from solar activity. ***VEPO is initially funded to assess the needs of the heliospheric research community, register highest interest data products for access through the VHO data query system, and work with the SPASE Consortium on improvement of data query terminology for application to VEPO-related data products. A longer-range goal is to support valued-added upgrades of selected data products for greater commonality of measurement parameters across multiple instruments on the same and different spacecraft with contiguous energy response ranges.*** The cross-calibration objective will enable improved applications of such data in the form of particle flux distributions over all relevant energies for study of interactions with geospace and other planetary environments, and for radiation hazard assessments of robotic and manned missions operating in interplanetary space. This effort will involve VEPO coordination and some proposed Data Upgrade projects when more funding is needed. Since energetic particles flow over wide energy ranges between the solar, heliospheric, planetary, and local interstellar environments, the scope of VEPO group activity could naturally extend beyond VHO to the other regional heliophysics VxOs, and even further to connections with earth and planetary science, with exploration in reference to radiation hazards, and with astrophysics in reference to external interstellar sources of cosmic rays. That is, VEPO is an example of a measurement-type focus group that could be crosscutting with respect to the regional VxOs supported within HDMC. ***During the NRA phase, VEPO will provide descriptions and access for Energetic Particle instruments on ACE, IMP-8, Voyager, Ulysses, Helios, New Horizons, Wind, Neutron Monitors, Pioneer, SOHO, and STEREO/IMPACT. As possible, but likely in an extended phase, metadata for Cassini, Galileo, Messenger, and STEREO/PLASTIC will be added.***

Newest: VWO, VMR

The last two of the VxOs were added in the most recent round of the NRA competition. Like VEPO, these cross disciplines, but fill in significant gaps in the coverage provided by the original VxOs.

Since wave phenomena are common occurrences throughout the Heliosphere, the primary objective of the Virtual Wave Observatory (VWO) is to provide basic wave data and information services to facilitate wave research in all Heliophysics domains: Sun, Heliosphere, Magnetosphere, Ionosphere, Atmosphere, and Planetary Magnetospheres. Like other VxOs, the VWO will: (1) work with the SPASE Group Consortium to develop the SPASE Data Model and provide wave data terms to the SPASE Data Dictionary; (2) describe the metadata of Heliophysics (space-based and ground-based) wave data sets using SPASE; (3) develop a Heliophysics wave data registry; and (4) develop a web interface for searching, subsetting and retrieving distributed wave data. The VWO differs from other domain-oriented VxOs in that the defining theme for VWO is the common interest in Heliophysics wave phenomena, irrespective of domains. The VWO thus aims to promote and facilitate interdisciplinary research that can lead to new and deeper understanding of wave processes and their relations to the structures and dynamics of various Heliophysics domains. To that end, the VWO goal is to make all Heliophysics wave data searchable, understandable and usable by the Heliophysics community. Since most wave data are recorded in spectral domain (e.g., frequency) in addition to time domain, and that wave data taken at a detector location (and time) could either be generated locally or from a remote source, searching and selecting wave data by time is not always the most meaningful route. As wave activities are often tied to the context conditions, such as solar or magnetospheric activity levels, associated with the wave sources, the VWO will develop context data search capability such that wave (and other Heliophysics) data can be searched by solar, solar wind and magnetospheric state conditions, as well as by time and location.

Wave data analysis often requires specialized knowledge of wave phenomena and the associated physics. This specialized expertise requirement can be a hindrance to wider use of Heliophysics wave data by non-wave researchers and students, despite the important information wave data may have for understanding many heliophysical processes. In addition to the basic data and information services, ***the VWO will enhance the understandability and usability of Heliophysics wave data by developing data annotation and tutorial services for***

describing the data content and illustrating how wave data can be used. Data annotations by expert users will be collected and organized into a searchable database, so that wave data can eventually be searched for specific features and for cross comparisons with other Heliophysics data sets. The innovative context and content data search capabilities will make VWO a very useful tool for Heliophysics research. ***The VWO plans to complete the descriptions of wave related datasets from IMAGE, Cluster, Polar, Geotail, and THEMIS during their NRA grant phase. To be included as possible during that time, or as necessary later, will be wave products from STEREO, Wind, Galileo, Cassini, DE, CRESS, Hawkeye, ISEE 1-3, Alouette-ISIS, Voyager, and Ulysses. In addition, a number of ground-based ULF, VLF, and radar datasets will be included.***

Global models are becoming more and more relevant in conducting Heliophysics research. ***The Virtual Model Repository seeks to make model data and runs as simply used as observational data in the course of Heliophysics research.*** Existing models span all regions of the geospace environment. As modeling grows, it becomes extremely important to validate the models by comparing them to data. This data can be derived from NASA missions or other sources. The main goal of the VMR is to make model output available to the community in a way that enables the direct comparison between data and models. An illustration of the approach was given in the Science Goals section above. The VMR is fully integrating into the VxO environment, using SPASE and extending it as needed (e.g., with grid and code information) to describe datasets, and integrating VxO access to observations to perform data-model comparisons. VMR plans include allowing users to search for both realistic and ideal model runs tied to data; creating an interface to NASA's ModelWeb; creating linkages between the VMR and other VxOs to allow models and data to be linked; and enabling contextual and comparative visualization of data within model results. ***VMR will work with model repositories located at the Community Coordinated Modeling Center, the University of Michigan, and the National Geophysical Data Center to start, using the Kameleon software from the CCMC to modify the data files at UM and NOAA to be uniform. In addition, VMR will support dynamic running of models from NASA's ModelWeb site.*** As the work progresses, new models must be added to the VxO environment. In addition, new data sets must be interpreted in terms of model results (or vice-versa). Depending on the complexity of the model or the data type, the task could be extremely complex or trivial, but the plan is to eventually incorporate models from all the subfields of Heliophysics. For an "average" case this task could take a programmer a month for each new data product or model, but this should become much easier as we gain experience.

Soon: SuperMAG

SuperMAG is a global collaboration that provides ground magnetic field perturbations from a long list of stations in the same coordinate system, identical time resolution and with a common baseline removal approach. This unique high quality dataset provides a continuous and nearly global monitoring of the ground magnetic field perturbation. The SuperMAG interface allows the user to choose data by time interval and location, making it easy to see what data are available and to preview them. Currently, only space born auroral imaging and the SuperDARN network of HF radars (when receiving backscatter) provide similar coverage of the auroral ionosphere. The primary support for SuperMAG has been from NSF, but ***with the addition of HDMC funding, SuperMAG will add auroral imaging obtained from a list of sources to the information from the ground based magnetometers, making this VxO-like service a uniquely valuable asset for users both directly and through VxOs.*** The use of the auroral emissions as a reference system for studies of the auroral electrodynamics has been proven to minimize the smearing of key auroral characteristics. The HDMC funding for SuperMAG was delayed while its basic functions were implemented. Starting in FY10, HDMC funding will allow SuperMAG to include auroral images from a list of imaging sources and to link to existing virtual observatories thereby expanding the capabilities of both SuperMAG and the virtual observatories. The provision of a VxO route to auroral imaging is, in itself, a significant addition to the VxOs.

Related: VSTO, GAIA, CSSDC, HELIO

While NASA currently has the most concerted program to develop virtual observatories in Heliophysics, there are many significant non-NASA efforts that will complement ours. These range widely in scope and purpose. The Canadian-based but internationally supported GAIA project unites ground-based all-sky cameras, riometers, and other data sources; they have participated in many VxO/HDMC meetings and are interested in interoperability with HDMC. The Virtual Solar-Terrestrial Observatory, primarily supported through NSF, unites access to the "CEDAR" database and the Mauna Loa Solar Observatory data. VSTO has also been active in HDMC activities. The University of Alberta is actively working on a ground-based data portal (CSSDC) that will be seamlessly integrated with the VMO. Finally, the HDMC is a collaborator on a European project very similar in scope to the HDMC called HELIO that will be funded by the European Commission under Framework Programme 7. This project has its roots in the European Grid of Solar Observations, a Virtual Observatory focusing on solar data. The

HELIO project is a multi-national, coordinated effort to provide uniform access to all Heliophysics data. It complements the HDMC in that it emphasizes aspects such as the exploitation of Event and Feature lists that have been less central to our efforts. We plan to work with HELIO to avoid duplication of effort and to agree on standards for Data Models and communication between systems and services, and to this end HDMC participated in the HELIO Kick-off meeting on 8-9 June 2009 in Paris. Some efforts have been made to collaborate with groups in other countries, such as Japan, and while there has been interest, these interactions are not as far along as those mentioned above.

Outreach, Evaluation, and Feedback

A critical component of any project such as the HDMC is obtaining community support. Anyone who has tried to develop tools of general utility knows of the gulf between having a valuable service and having it used by anyone but the developers. There is also an inherent tension between releasing things so that people can try them, and holding back to make sure that the trials will not lead to rejection due to the problems encountered. Our approach to this will be to continually upgrade our services in a way that maintains reliability but adds something useful with each change. We will assign high priority to any problems found by users that are of general importance to utility or robustness.

Once we attain a basic level of utility for users, we will advertise widely. In addition to making continual announcements in the general newsletters and publications of our organizations, we will continue to establish connections with missions and users, directly and through meetings. We will have a booth or share a NASA booth at the AGU and other meetings as possible. Something we hope will be very effective is to make better use of modern Internet capabilities. Instead of (or in addition to) having long pages of help text, short videos and screencasts will explain features of the systems. We will use direct links to short videos in the subject lines of newsletters to add interest. We will use humor and graphics to catch people's attention. We will use modern tools to create more effective Web interfaces. Critical to all this will be to have good services, worth advertising. The effort of making effective "sales pitches" may well lead us to see shortcomings in our systems.

We will continually seek feedback from different members of the community. Most formally, this will come from HQ and the HPDCWG. Less formally, the Project Scientist and each team will specifically ask individuals and user groups to try the services and provide feedback on a regular basis. This will both improve the services and introduce more people to them.

The formal evaluation of any data system is difficult to quantify. Web hits are often used, and certainly few hits would be a sign of failure and many hits are good, but these are not a very reliable indicator of utility. While it is often said that data sources are not often acknowledged in papers, the simple presence of the name of the spacecraft or observatory and instrument go a long way to indicating the source. (The references to data and the acknowledgments of sources will be made much easier by the Heliophysics Inventory.) Repositories and services can much more easily be overlooked, and thus it may be difficult to ascertain how many papers relied on a VxO or a related service in performing research. While we plan to track publications and to encourage the acknowledgment of services in papers, we believe that formal and informal community feedback, the support of missions, and progressively more sophisticated automated accounting of the actual usage of the VxOs and other services will be the most effective ways of judging success. If we are completely successful, the system we are proposing will simply be a routine part of the way scientists obtain and interact with data, as clearly useful as abstract services and journal articles.

Plan of Work and Milestones

Management Plan

The management plan is simple, and was given in the introduction. The Project Scientist has the overall responsibility for the success of the project, and will also handle funding issues, the planning of meetings, the preparation of Senior Review proposals, etc. Representatives of the components of the HDMC and the other Data and Modeling Centers form the Implementation Working Group that determines the direction of the project, and the PIs of each component are responsible for carrying out the work. Review and feedback are provided by the HPDCWG; component-by-component oversight and user groups; and the Data Center Senior Reviews.

Milestones

Inventory and SPASE descriptions (Task 1, 6)

Within a year, complete an initial (active) inventory of all Heliophysics resources, and complete and implement tools for SPASE-based resource registry management.

Within two years, complete at least basic SPASE descriptions and proper final ID assignments/registration of all Heliophysics resources. (At this point, the data product ID will be useful as a reference in published papers, as well as in general tools, e.g., for referring to a data reader format or snippet of IDL code, or for access by a service such as VSPO or any VxO: ID + time range, perhaps + a variable list yields a data result.)

Discipline-Specific Uniform Data Discovery and Access (Task 2; progress is VxO dependent)

Within the first year, improve Web access through VxO portals to an initial set of data products. Share insights into what works best between VxOs. Systematically exploit user groups to correct and refine approaches.

Within two years, have fully functioning portals in all areas. Within three years, implement the main value added services promised for each portal and implement sharing of such services wherever possible.

Within three years, optimize the type and set of interfaces to accommodate different user styles and remove redundancy.

Browsing and Visualization (Task 3)

Within a year, produce a stable tool for reading and browsing the most common format of Heliophysics data that includes basic time-series plotting capabilities.

Within two years, extend the basic service to include spectrograms, and optimize the interface to deal with multiple datasets.

Within three years, provide two- and three-dimensional capabilities, linking them to VxOs and the active Inventory.

Within three years, use Autoplot and other routes to provide a service that can be called from, e.g., IDL or Java tools to obtain data in the desired form. Incorporate search tools as well.

Dataset Runs on Request/Download Service (Task 4)

Within one year, provide a Dataset Runs-on-Request service that work for a limited set of formats (e.g., CDF), but that allows subsetting, merging, interpolating, averaging, choosing of missing data flags, and merging data from different instruments or observatories.

Within two years, eliminate the “formats problem” by providing a stable tool for reading and producing a variety of formats (ASCII, IDL saveset, MatLab .mat, CDF, HDF, etc.). Increasingly provide dataset utility at a semantic level.

Within three years, develop Dataset Runs on Request that uses the “formats solution” to ingest any desired datasets and produce desired output, and fully populate the required database of simple accessors to enable the solution to work.

VxO access; SPASE QL (Task 7)

Within a year, provide easy, one-stop browser and API access to data within subfields that exploits the SPASE categorization and returns data to the user for all of the most commonly used datasets; provide links between the subfields to allow cross-domain access.

Within two years, provide the above service for a comprehensive set of data.

Resident Archives (Task 8)

Within the first year, formalize the oversight of Resident Archives to ensure proper data stewardship of Heliophysics resources with a minimum of management infrastructure.

Within the second and subsequent years, continue to evolve the RAs, developing clearer criteria for their continuation. Add RAs for missions as they end through the NRA process.

Data Upgrades and Value Added Services (Tasks 9 and 5)

Continue NRA calls to maintain community review and involvement in providing both Data Upgrades and Value Added Services. Evaluate each year the community response and needs to set the funding level for these appropriately *vis-à-vis* other uses for the funding.

Years 4 and 5 (all Tasks)

Maintain the above infrastructure for VxOs and data services. Continue to evolve Resident Archives. Continue Data Upgrades as necessary. Continue to populate and improve data descriptions to make resources more useful; keep up with new missions and products. Do a regular systematic review of resources and Inventory for uniformity, correctness, and completeness. Continue to define and evolve services, largely through NRA competition. Do a systematic review of archives.

Throughout all developments, work with national and international partners (NSF, NOAA, DoD; HELIO, CSSDP, GAIA, etc.) to ensure the development of standards for interoperability that will make the Heliophysics Data Environment as broadly comprehensive as it should be.

Budget Details and Justification

In-Guide

Appendix I provides a breakdown of the HDMC budget by functions. It is difficult to predict exactly what the breakdown of the budget will be because it depends on, for example, how much funding for RAs will be needed when missions end, and this and other elements of the picture are not predictable. Overall, however, we expect that the RA portion of the budget will grow to about 25% of the total and then remain relatively fixed as older RAs are no longer needed, making up for new ones. We are budgeting a fixed Data Upgrade line, at roughly 10% of the total, but this depends on NRA demand for these services. The budget for new annual NRA funding includes the Data Upgrades, new RAs, and new or upgraded Value Added Services; we plan on keeping this a fixed fraction of the total at about 20%, but this could change somewhat as needs are evaluated. Project Scientist support will require 5% of the budget, perhaps declining somewhat as operations become more routine.

Initially, nearly half (45%) the budget goes to VxOs directly or to support the core services such as visualization (Autoplot) and Data Services (Runs-on-Request/DataShop). This funding will be half devoted to finishing the inventory and descriptions of data products, with the rest devoted to interface development, community interactions, interoperability (SPASE QL, etc.), and registry services. As the registries become well populated with products, the emphasis will shift to providing easy access to core services and to enhancing, for example the utility of event lists and other advanced query methods.

The budget of \$700K/yr for two years to complete the descriptions of a comprehensive set of resources, and thus make them available through the VxOs, may seem excessive. However, this is much less than 1 FTE/year for each VxO, and most of the VxOs have a substantial number of products to be described. Generally this process can be at best partially automated. The experience of many Data Centers shows that data providers are reluctant to spend any of their limited resources on such tasks, and thus it is up to VxO personnel, or in some instances a provider who receives specific funding, to do the task. The people needed to create these descriptions are scientists, not more affordable support personnel. This is largely a one-time expense, and the experience gained to date has made the process much more efficient. As a calibration for this expense, the Cluster Active Archive has spent more very detailed descriptions for one spacecraft than we are asking. The trick is to choose the right level of detail for our purposes such that we will be able to accomplish the task in a reasonable time, and we now know how to do this. (For future missions, the task of making complete SPASE descriptions will be a very small fraction of the missions' data system budgets, and this will be planned in from the start as a mission and/or VxO function.)

Similar considerations are relevant to the other categories of VxO activity. With the In-Guide budget, each continuing VxO will receive enough to cover ~1 FTE, which will have to be spread between the descriptions of resources, development of interfaces, and some development of services. The service development, while not

initially primary, is essential to provide the unique functionality that will show the community that the approach is worthwhile. Unless we receive over-guide funding, there is no way to shift resources into meeting the initial requirements. The expense associated with VxOs will decrease with time as the core of the systems becomes established, but there will be an irreducible minimum to maintain service levels.

The rest of the budget covers the RAs, Data Upgrades, and Value Added Services that will be evaluated in NRA review. The main issue is how much of the budget should be devoted to these activities as opposed to VxO activities. The RAs are the least flexible of the requirements, although difficult to predict since it depends on the uncertain times when missions will end. In the initial years, the requirements of the VxOs force the NRA budget to be smaller. The trade-offs presented in the tables give our current best estimates, to be reevaluated in Program budget reviews starting next year for the FY12 budget.

Optimal

For the optimal budget, we are requesting an additional \$500K per year for two purposes: jump-starting the VxOs and providing more capable services in the out years. An extra \$300K/year for the first two years would make it possible to provide a much better set of resource descriptions, including both more complete descriptions and documentation of NASA resources, and a more complete extension to non-NASA resources. The other \$200K/year for the first two years would be used to more quickly produce browsing and Runs-on-Request tools, including the required accessor descriptions. Neither of these tasks is qualitatively different from what is described in the plans above, but more resources would allow the task to be done more quickly and completely. This would be of very great benefit to the program, since it would increase user buy-in in the early stages. Thus we would more quickly solve the two central problems of completeness and format independence. For the additional years, increased funding would allow the more complete development of, for example, event-based and data mining tools and their deeper integration into the VxO framework. These are the sorts of services that the Virtual Astronomical Observatory (formerly NVO) are developing as essential to realizing the full promise of virtual observatories. This would greatly increase the utility of the system.

References (URLs and Acronyms)

Autoplot <http://autoplot.org>
DataShop <http://adsabs.harvard.edu/abs/2004AGUFMSH21B0418V>
HDMC Publication List <http://www.spase-group.org/biblio.jsp> and the references provided at many VxO sites.
Heliophysics Data Environment (HPDE; site contains many general references and the Heliophysics Science Data Management Policy) <http://hpde.gsfc.nasa.gov>
Heliophysics Integrated Observatory (HELIO) <http://www.helio-vo.eu/>
Heliophysics Resource Gateway (Virtual Space Physics Observatory; VSPO) <http://vspo.gsfc.nasa.gov>
National Space Science Data Center (NSSDC) <http://nssdc.gsfc.nasa.gov>
Space Physics Data Facility (SPDF) <http://spdf.gsfc.nasa.gov>
SuperMAG (worldwide collaboration of ground based magnetometers) <http://supermag.jhuapl.edu>
Virtual Energetic Particle Observatory (VEPO) <http://vepo.gsfc.nasa.gov>
Virtual Heliospheric Observatory (VHO) <http://vho.nasa.gov>
Virtual Ionosphere, Thermosphere, Mesosphere Observatory (VITMO) <http://vitmo.jhuapl.edu>
Virtual Radiation Belt Observatory (ViRBO) <http://virbo.org>
Virtual Magnetospheric Observatory (VMO) <http://vmo.nasa.gov> and <http://vmo.igpp.ucla.edu>
Virtual Model Repository (VMR) <http://adsabs.harvard.edu/abs/2008AGUFMSA53A1572D>
Virtual Solar Observatory (VSO) <http://virtualsolar.org>
Virtual Wave Observatory (VWO) <http://vwo.gsfc.nasa.gov>
Visual System for Browsing, Retrieval, and Analysis of Data (ViSBARD)
<http://spdf.gsfc.nasa.gov/research/visualization/visbard/>

Other Acronyms

ACE: Advanced Composition Explorer <http://www.srl.caltech.edu/ACE/>
Alouette I and II <http://www.asc-csa.gc.ca/eng/satellites/alouette.asp>
AP8/AE8: Integral Proton/Integral Electron Trapped Particle Flux Maps
API: Application Programming Interface
ASCII: American Standard Code for Information Interchange
CAA: Cluster Active Archive <http://caa.estec.esa.int>
CCMC: Community Coordinated Modeling Center <http://ccmc.gsfc.nasa.gov/>
CDAWeb: Coordinated Data Analysis Web <http://cdaweb.gsfc.nasa.gov/>
CDF: Common Data Format <<http://cdf.gsfc.nasa.gov/>>
Cluster <<http://caa.estec.esa.int/caa/>>
CME: Coronal Mass Ejections
C/NOFS: Communication and Navigation Outage Forecast System
<http://www.fas.org/spp/military/program/nssrm/initiatives/cnofs.htm>
COHOWeb <<http://cohoweb.gsfc.nasa.gov/>>
HPDCWG: Heliophysics Data and Computing Working Group
<<http://igsb.jpl.nasa.gov/mail/igs-dcwg/igs-dcwg.html>>
DE: Dynamics Explorers 1 and 2 <http://nssdc.gsfc.nasa.gov/nmc/masterCatalog.do?sc=1981-070A>
<http://nssdc.gsfc.nasa.gov/nmc/masterCatalog.do?sc=1981-070B>
FAST: Fast Auroral Snapshot Explorer <http://sprg.ssl.berkeley.edu/fast/>
FITS: Flexible Image Transport System <http://fits.gsfc.nasa.gov/>

FTP: File Transfer Protocol
 FTPBrowser: <http://ftpbrowser.gsfc.nasa.gov/>
 FY: Fiscal Year
 Geotail <http://pwg.gsfc.nasa.gov/geotail.shtml> <http://www.stp.isas.jaxa.jp/geotail/>
 GOES: Geostationary Operational Environmental Satellites <http://www.oso.noaa.gov/goes/>
 GSFC: NASA Goddard Space Flight Center <http://www.nasa.gov/centers/goddard/home/index.html>
 HDF: Hierarchical Data Format <http://hdf.ncsa.uiuc.edu/>
 HDMC: Heliophysics Data and Modeling Consortium <http://hpde.gsfc.nasa.gov/>
 HELM: Heliophysics Event List Manager <<http://helm.gsfc.nasa.gov>>
 HTTP: Hyper Text Transfer Protocol
 IBEX: Interstellar Boundary Explorer <http://www.ibex.swri.edu/>
 IDL: Interactive Data Language <http://www.itvis.com/>
 IMAGE: Imager for Magnetopause-to-Aurora Global Exploration <http://image.gsfc.nasa.gov/>
 IMPACT (STEREO): In-situ Measurements of Particles and CME Transients
 <<http://sprg.ssl.berkeley.edu/impact/>>
 IMP-8: Interplanetary Monitoring Platform <http://spdf.gsfc.nasa.gov/imp8/project.html>
 ISAS: Institute of Space and Astronautical Science, Japan
 ISEE-3: International Sun-Earth Explorer 3
<http://nssdc.gsfc.nasa.gov/nmc/masterCatalog.do?sc=1978-079A>
 ISIS: International Satellites for Ionospheric Studies
 ISTEP: International Solar Terrestrial Program <http://pwg.gsfc.nasa.gov/istp/>
 ITM: Ionosphere-Mesosphere-Thermosphere
 LANL: Los Alamos National Laboratory satellites <http://www.lanl.gov/>
 LWS: Living with a Star <<http://lws.gsfc.nasa.gov/>>
 NASA: National Aeronautics and Space Administration
 NOAA: National Ocean and Atmospheric Administration <http://www.noaa.gov/>
 NRA: NASA Research Announcement
 NSSDC: National Space Science Data Center <<http://nssdc.gsfc.nasa.gov/>>
 NSSDCftp <http://nssdcftp.gsfc.nasa.gov/>
 OMNIweb: <http://omniweb.gsfc.nasa.gov/>
 PDMP: Project Data Management Plan
 PLASTIC (STEREO): Plasma and Suprathermal Ion Composition <http://stereo.sr.unh.edu/>
 Polar satellite <http://pwg.gsfc.nasa.gov/polar/>
 RAs: Resident Archives
 RBSP: Radiation Belt Storm Probes <http://rbsp.jhuapl.edu/>
 REST: Representational State Transfer
 <http://en.wikipedia.org/wiki/Representational_State_Transfer>
 RHESSI: Reuven Ramaty High Energy Solar Spectroscopic Imager
<http://hesperia.gsfc.nasa.gov/hessi/>
 ROCSAT: Republic of China Satellites
 SDAC: Solar Data Analysis Center <http://umbra.gsfc.nasa.gov/sdac.html>
 SMWG: SPASE Metadata Working Group <http://www.spase-group.org/registry/>
 SNOE: Student Nitric Oxide Explorer <http://lasp.colorado.edu/snoe/>
 SOHO: Solar & Heliospheric Observatory <http://soho.nascom.nasa.gov/>
 SolarSoft <http://sohowww.nascom.nasa.gov/solarsoft/>
 SPASE: Space Physics Archive Search and Extract <http://www.spase-group.org/>
 SPASE QL: SPASE Query Language <http://vho.nasa.gov/vxo/spaseql.php>
 SPDF: Space Physics Data Facility <http://spdf.gsfc.nasa.gov>
 SSCWeb: Satellite Situation Center Web <http://sscweb.gsfc.nasa.gov/>
 STEREO: Solar TERrestrial RELations Observatory <http://stereo.gsfc.nasa.gov/>
 THEMIS: Time History of Events and Macroscale Interactions during Substorms:
<http://sprg.ssl.berkeley.edu/themis/flash.html>
 TIMED: Thermosphere-Ionosphere-Mesosphere Energetics and Dynamics
<http://www.timed.jhuapl.edu/WWW/index.php>
 TWINS: Two Wide-angle Imaging Neutral-atom Spectrometers <http://twins.swri.edu/index.jsp>
 ULF: Ultra Low Frequency

URL: Uniform Resource Locator
VEFI (DE and C/NOFS): Vector Electric Field Instrument
VLF: Very Low Frequency
VO: Virtual Observatory
Voyager 1,2 satellites <http://voyager.jpl.nasa.gov/>
VWO: Virtual Waves Observatory <http://vwo.gsfc.nasa.gov/>
VxOs: Virtual Discipline Observatories
XML: Extensible Markup Language